UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Máster en Análisis Avanzado de Datos Multivariantes

Trabajo Fin de Máster

ESTUDIO DE LA PRECISIÓN DE LOS RESULTADOS DEL HJ-BIPLOT A TRAVÉS DE LOS MÉTODOS BOOTSTRAP

AUTOR: Ana Belén Nieto Librero

TUTOR: Purificación Galindo Villardón

Año 2013

Índice

Resumen

Los métodos Biplot son técnicas de análisis multivariante. Un biplot es una representación gráfica de datos multivariantes, donde se pretende aproximar los elementos de una matriz a partir de marcadores para las filas y columnas de la misma con el fin de representarlos en un espacio de dimensión reducida. Estos marcadores nos permiten calcular más parámetros que nos ayudan a la interpretación de los resultados tales como: bondades de ajuste, calidades de representación, contribuciones de los elementos y los los ejes, variabilidad y relaciones entre variables… Sin embargo, todos estos parámetros son estimaciones puntuales obtenidas a partir de una muestra de datos y no proporcionan ninguna información acerca de la incertidumbre proporcionada por estas medidas.

En este trabajo se revisan los principales aspectos de los métodos Biplot así como la manera de interpretarlos y se presentan las ideas básicas de los métodos bootstrap con el objetivo de utilizarlos para proporcionar una forma de calcular medidas de incertidumbre para la información que presentan los métodos Biplot. La manera propuesta de medir la precisión es a través del cálculo de intervalos de confianza basados en percentiles e intervalos t-bootstrap. Dichos intervalos se calculan a partir de los resultados obtenidos mediante remuestreo bootstrap. Para ello se ha desarrollado un paquete en el entorno R. Como ejemplo se presentan los resultados obtenidos al realizar un análisis HJ-Biplot en un conjunto de datos reales y en un conjunto de datos simulado y se calculan intervalos de confianza de sus parámetros mediante la metodología bootstrap.

Palabras clave: HJ-Biplot, bootstrap, intervalos de confianza, percentiles.

Introducción

Los métodos de ordenación y de reducción de la dimensión ofrecen diferentes parámetros para poder presentar los resultados obtenidos a partir de una muestra de datos multivariante. Sin embargo, estos métodos únicamente nos muestran esos resultados de una manera incompleta ya que sólo se obtienen estimaciones de tales parámetros de una manera puntual, sin ninguna información acerca de la incertidumbre proporcionada por los mismos. Para proporcionar unos resultados completos es necesario mostrar una forma de decidir cómo de exactos son esos parámetros. El método más común para proporcionar una indicación de la cantidad de incertidumbre de un parámetro son los intervalos de confianza representados por los límites de confianza.

(Davison & Hinkley, 1997; Efron & Tibshirani, 1993; Efron, 1979, 1987) introdujeron los métodos bootstrap, cuya idea básica es que la inferencia sobre una población a partir de una muestra se puede obtener remuestreando dicha muestra y haciendo inferencia sobre esta nueva "muestra". Estos métodos proporcionan diferentes formas de calcular intervalos de confianza para los parámetros calculados a partir de una muestra de datos multivariantes. Tienen la ventaja además de que son métodos sencillos que no requieren del conocimiento de la distribución teórica de la población de partida y tampoco necesitan un tamaño de muestra elevado para realizar las estimaciones.

Estos métodos han sido utilizados combinados con diversas técnicas multivariantes para proporcionar resultados más precisos. (Gifi, 1990; Greenacre, 1984; Meulman, 1982) fueron los que introdujeron estos métodos en el contexto de las dos vías o el Análisis de Correspondencias Múltiple (MCA); (Chatterjee, 1984; Lambert, Wildt, & Durand, 1990, 1991) lo utiliza en el contexto del Análisis Factorial; en el caso del Análisis de Componentes Principales (PCA) (Daudin, Duby, & Trécourt, 1988; Diaconis & Efron, 1983; Holmes, 1985, 1989; Stauffer, Garton, & Steinhorst, 1985) utilizaron la metodología bootstrap para proponer intervalos de confianza para los puntos representados en el subespacio de los ejes principales intentando resolver el problema de la elección del número de ejes a retener; (Milan & Whittaker, 1995) lo proponen en el caso de modelos bilineales que incorporan Descomposición en Valores Singulares (SVD); (Raykov & Little, 1999) se valen de los métodos bootstrap para evaluar el ajuste de las rotaciones Procrustes;

(Linting, Meulman, Groenen, & Van der Kooij, 2007) en el PCA no lineal; (Kiers, 2004) lo utiliza para los resultados de métodos de tres vías; (Lavoranti, dos Santos, & Kraznowski, 2007) analiza la estabilidad fenotípica a través de modelos AMMI con remuestreo bootstrap; (Timmerman, Kiers, Smilde, Ceulemans, & Stouten, 2009) utiliza los métodos bootstrap para estimar intervalos de confianza en el Análisis de Componentes Multinivel (MLSCA); (Van Ginkel & Kiers, 2011) proponen una forma de corregir los resultados del bootstrap en el PCA en caso de presencia de valores perdidos.

Siguiendo este procedimiento, se propone la combinación del remuestreo bootstrap con el análisis HJ-Biplot. De esta manera, se ofrecen resultados completos mediante el cálculo de intervalos de confianza basados en percentiles e intervalos t-bootstrap de los parámetros proporcionados por el análisis. Para facilitar el uso de esta combinación de análisis se ha desarrollado un paquete en el entorno R llamado *biplotbootGUI* que permite análisis biplot clásicos (HJ, GH y JK) con la ventaja de poder realizar sobre los datos un remuestreo bootstrap y obtener intervalos de confianza para las principales medidas obtenidas por dichos análisis.

En el apartado Material y Métodos se explica la estructura de los dos conjuntos de datos que se han analizado. A continuación, se resumen los principales aspectos de los métodos Biplot, de la metodología bootstrap y del lenguaje R. Por último se presentan los diferentes paquetes existentes en R que realizan algún tipo de biplot y aquéllos que facilitan resultados de la metodología bootstrap y se muestra el nuevo paquete implementado.

En el apartado Resultados se realizan análisis HJ-Biplot de los dos conjuntos de datos utilizados y se muestran los histogramas y los gráficos de normalidad en los que se representan los conjuntos de valores obtenidos para cada medida que se calcula en el HJ-Biplot así como los intervalos de confianza basados en percentiles y t-bootstrap para cada uno de ellos.

Objetivos

El objetivo general del presente trabajo es ofrecer una nueva forma de presentar los resultados de un análisis HJ-Biplot en el que se puedan calcular no sólo las medidas puntuales que proporciona dicho análisis tales como: bondad de ajuste, calidad de representación de filas y columnas, valores propios y variabilidad y relación entre variables; sino que también permita construir intervalos de confianza para cada uno de ellas basados en los percentiles e intervalos t-bootstrap que se pueden obtener a partir de un remuestreo bootstrap para poder medir su precisión.

Como objetivos específicos se tienen:

- Hacer una breve descripción de los métodos Biplot, la información que proporcionan así como la manera de interpretarla.
- Presentar un resumen de los principales aspectos de la metodología bootstrap, y sus diferentes formas de calcular intervalos de confianza.
- Desarrollar una herramienta en el entorno R que permita la utilización de esta propuesta de una manera fácil, rápida y flexible.
- Mostrar los resultados de esta propuesta aplicado a un conjunto de datos reales y a un conjunto de datos simulados.

Material y Métodos

Se analizaron dos conjuntos de datos. El primero, llamado iris (Fisher, 1936), corresponde al famoso conjunto de datos recogidos por Anderson en 1935 (Anderson, 1935) y que constan de las medidas en centímetros de las variables longitud y anchura del sépalo y longitud y anchura del pétalo de 50 flores provenientes de 3 especies de iris. Las especies son *Iris setosa*, *versicolor* y *virginica*. El segundo es una matriz de datos simulada que consta de 100 medidas de 5 variables que siguen una distribución normal y que han sido elegidas convenientemente para que tengan las siguientes correlaciones entre sí:

- Correlación(V_1, V_3)=0.80
- Correlación(V_2, V_3)=0.50
- Correlación(V_4, V_5)=0.90

Biplot

Introducción

Los métodos Biplot (Gabriel, 1971; Galindo & Cuadras, 1986; Galindo, 1986) están basados en la descomposición en valores singulares de una matriz al igual que otras técnicas factoriales para obtener una representación de la información en un espacio de baja dimensión con la menor pérdida de información posible. Permiten el análisis de una tabla de dos vías que contiene información de J variables medidas sobre I individuos.

El prefijo "bi" se refiere a la representación simultánea tanto de filas como de columnas de la matriz de partida lo cual es una ventaja respecto a otras técnicas.

Las representaciones Biplot están basadas en las propiedades geométricas del producto escalar entre marcadores fila y marcadores columna de tal forma que cada elemento de la matriz es aproximado por este producto.

Marco teórico

Sea X una matriz de datos de rango $\tilde{S}$ que contiene información de I individuos medidos en J variables y sea $I > J$ es decir, $\min(I, J) = J$.

Un Biplot para la matriz de datos X es una representación gráfica mediante marcadores $a_1, a_2, ..., a_I$ para las filas de X y $b_1, b_2, ..., b_J$ para las columnas de X, de tal forma el producto interno aproxime el elemento x_{ij} de la matriz de partida lo mejor posible (Figura 2).

Si consideramos los marcadores $a_1, a_2, ..., a_I$ como filas de una matriz A y los marcadores $b_1, b_2, ..., b_J$ como filas de una matriz B, entonces se puede expresar la matriz X de partida como:

$$X = AB^T .$$

Si el rango de la matriz X es ≤ 3 es posible obtener una representación gráfica sin pérdida de información para el biplot en una recta, un plano o un espacio tridimensional respectivamente.

Si el rango de la matriz X es > 3 no es posible obtener una representación gráfica sin pérdida de información.

La descomposición en valores singulares de la matriz X se define como:

$$X = \mathrm{U}\Lambda\mathrm{V}^{\mathrm{T}}$$

Donde:

U es una matriz cuyos vectores columna son ortonormales y son los vectores propios de la matriz XX^T.

V es una matriz cuyos vectores columna son ortonormales y son los vectores propios de la matriz $X^T X$.

Λ es una matriz diagonal que contiene los valores singulares de la matriz X, que son las raíces cuadradas no negativas de los valores propios de $X^T X$, ordenados en forma decreciente.

Para que las matrices U y V sean ortonormales debe cumplirse que $U^T U = V^T V = I$. Esta propiedad asegura la unicidad de la factorización.

Un elemento genérico de la matriz X puede ser escrito como:

$$x_{ij} = \sum_{s=1}^{\min(I,J)} \lambda_s u_{is} v_{js} \qquad (0.1)$$

Para encontrar la aproximación en un espacio de baja dimensión de la matriz X es necesario minimizar la distancia entre la matriz original $X = x_{ij}$ y la matriz aproximada $\tilde{X} = \tilde{x}_{ij}$. Esta distancia (Euclídea) entre dos matrices se define como:

$$d(X, \tilde{X}) = \sqrt{\sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \tilde{x}_{ij})^2}$$

El teorema de Eckart y Young (Eckart & Young, 1936), que también puede encontrarse en otros autores como (Gabriel, 1971; Greenacre, 1984; Young & Householder, 1938) demuestra que la mejor aproximación S -dimensional de la matriz X en el sentido de los mínimos cuadrados se puede obtener mediante la descomposición en valores singulares de la matriz X sumando únicamente los S primeros términos en la ecuación (0.1).

Los primeros S u_s y v_s combinados con los valores singulares λ_s de diferentes maneras constituyen la forma de calcular las coordenadas de los datos para su representación gráfica. En el caso de la métrica identidad, se pueden calcular de acuerdo a los diferentes valores de en la siguiente descomposición:

$$X = AB^T$$

$$\mathrm{A} = \mathrm{U}\Lambda^{\gamma}$$

$$\mathrm{B} = \mathrm{V}\Lambda^{1-\gamma}$$

Dependiendo del valor que tome γ ($\gamma = 0, 1, 1/2$) se obtienen los Biplots Clásicos de Gabriel (Gabriel, 1971) GH-Biplot y JK-Biplot y el SQRT-Biplot respectivamente.

Los más utilizados se muestran en la Figura 1.

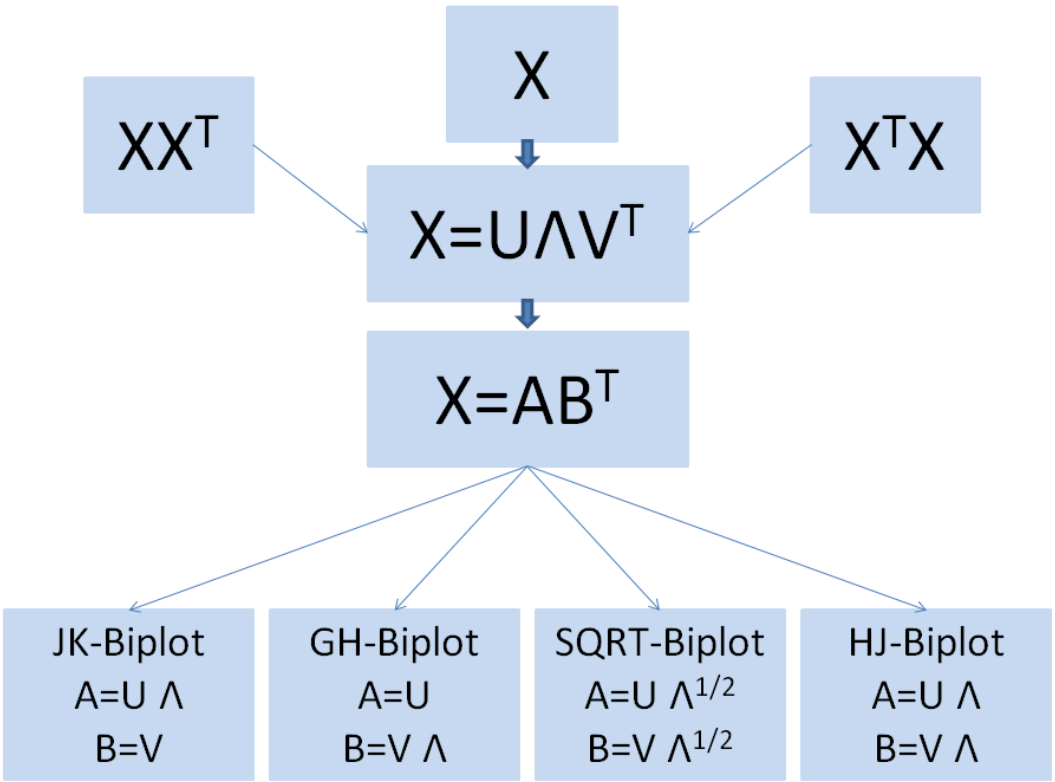


Figura 1. Biplots.

Interpretación Geométrica de las representaciones Biplot

Un biplot estándar es la representación de un elemento de la matriz (interacción individuo por variable). La matriz X se descompone como un producto AB^T, donde A es una matriz de dimensión IxS y B es una matriz de dimensión JxS. Utilizando la descomposición en dimensión S de la matriz aproximada, cada elemento se puede escribir como:

$$\tilde{x}_{ij} = \sum_{s=1}^{s} a_{is} b_{js}$$

que es el producto escalar de los vectores $\left(a_{i1}, a_{i2}, ..., a_{iS}\right)$ y $\left(b_{j1}, b_{j2}, ..., b_{jS}\right)$ (Figura 2).

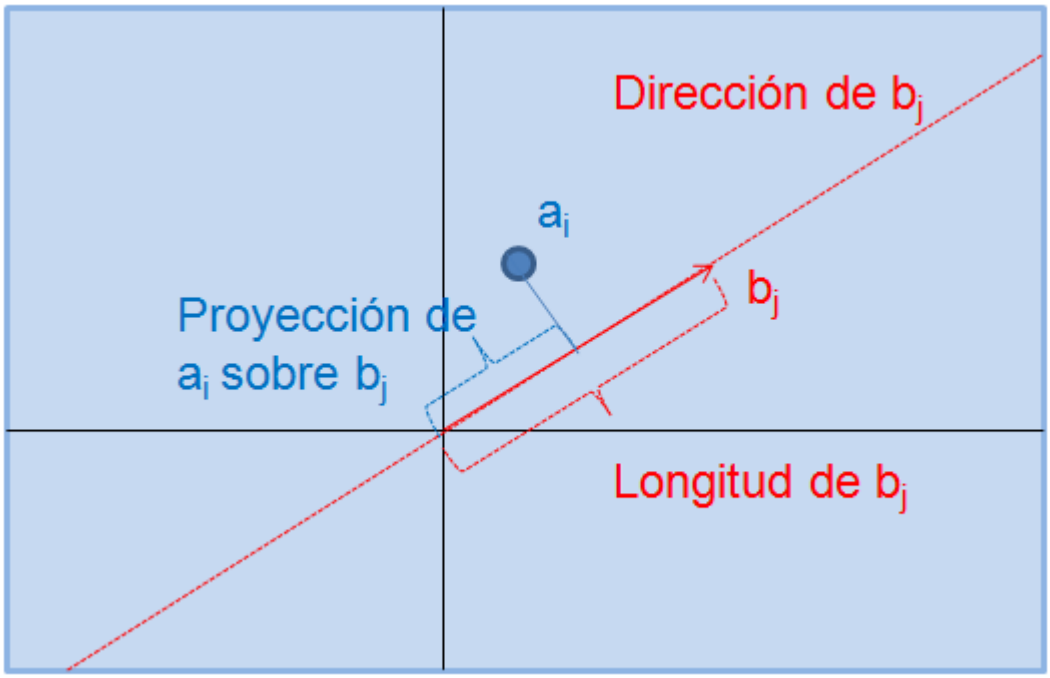


Figura 2. Producto Escalar.

Por tanto, un Biplot se obtiene representando cada fila como un punto con coordenadas $\left(a_{iS}\right)$ y cada columna como un vector con coordenadas $\left(b_{jS}\right)$ con $s = 1, ..., S$ en un espacio Euclídeo de dimensión S respecto a los mismos ejes ortogonales. Estos puntos generalmente son referidos como marcadores fila y marcadores columna. Este tipo de representación facilita la proyección de los marcadores fila sobre los marcadores columna (Figura 3).

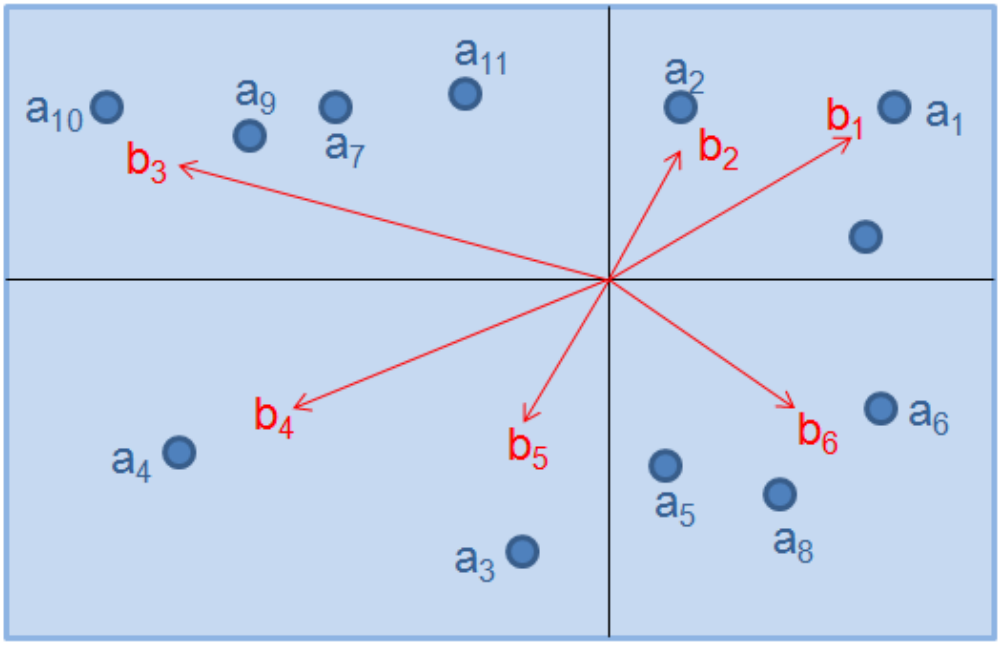


Figura 3. Ejemplo de representación Biplot.

Las relaciones entre individuos y variables son estudiadas a través de las proyecciones de los puntos que representan a los individuos sobre los vectores que representan a las variables. Es decir,

$$x_{ij} \approx a_i^T b_j \Rightarrow x_{ij} \approx \left\| proy\, a_i / b_j \right\| signo \left\| b_j \right\|$$

donde:

$\left\| proy\, a_i / b_j \right\|$ = longitud del segmento que va desde el origen de la representación hasta el punto a_i (longitud de la proyección de a_i sobre b_j).

$\left\| b_j \right\|$ = módulo de b_j (longitud del segmento que une el origen con el extremo del vector b_j).

O sea, x_{ij} es aproximadamente el módulo de la proyección de a_i sobre b_j multiplicado por la longitud de b_j , con el signo correspondiente.

La dirección del vector b_j muestra la dirección en la que aumentan los valores de la correspondiente variable. Las proyecciones de los puntos a_i sobre un vector columna aproximan la j -ésima columna de la matriz X , y proporciona un orden de los individuos respecto de dicha variable.

Una vez definida la forma de representación, podemos interpretar:

- La distancia entre individuos como disimilaridades entre ellos, es decir, una menor distancia entre individuos implica una menor disimilaridad entre ellos, especialmente si los individuos están bien representados.

- Las longitudes y ángulos de los vectores que representan a las variables como variabilidad y covariabilidad respectivamente en el GH-Biplot.
- Las relaciones entre individuos y variables en términos de productos escalares, es decir, a través de las proyecciones de los puntos que representan individuos sobre los vectores que representan a las variables.
- El orden de los individuos respecto de cada variable a través de la proyección de los puntos individuos sobre el eje cuya dirección está marcada por el vector que representa cada variable.

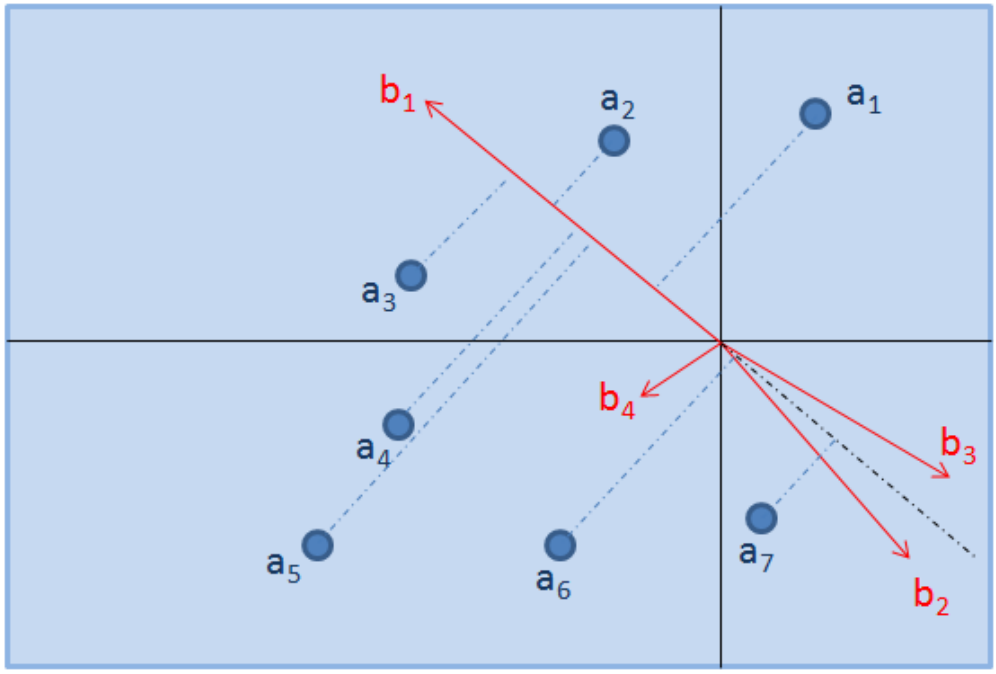


Figura 4. Representación de una matriz 7x4.

Estas interpretaciones se muestran en la Figura 4. En ella se puede ver que la variable 1 y la variable 4 son independientes (ya que el ángulo que forman los dos vectores es aproximadamente 90). Las variables 1 y 2 tienen una alta correlación negativa (el ángulo que forman está próximo a 180). Las variables 2 y 3 tienen una alta correlación positiva (el ángulo que forman es muy pequeño).

La variable 4 es la variable que presenta menor variabilidad y la variable 1 es la que presenta mayor variabilidad debido a que sus vectores son los de menor y mayor longitud respectivamente.

El orden decreciente de los individuos para la variable 1 es 3, 2, 4, 5, 1, 6 y 7 como se puede apreciar al proyectar los puntos que los representan sobre la dirección de dicha variable.

En las representaciones Biplot también es posible mostrar combinaciones lineales de filas y columnas. Ejemplos de estas combinaciones son las medias de filas y de columnas.

- Las medias de las filas estarán ordenadas como las proyecciones de los marcadores fila sobre el vector que representa al marcador columna medio, puesto que:

$$x_{\bullet j} \approx a_{\bullet}^{T} b_j \approx \left\| proy\, b_j / a_{\bullet} \right\| signo \left\| a_{\bullet} \right\| = \left\| proy\, a_{\bullet} / b_j \right\| signo \left\| b_j \right\|.$$

- Las medias de las columnas estarán ordenadas como las proyecciones de los marcadores columna sobre el vector que representa al marcador fila medio, ya que:

$$x_{i\bullet} \approx a_i^{T} b_{\bullet} \approx \left\| proy\, b_{\bullet} / a_i \right\| signo \left\| a_i \right\| = \left\| proy\, a_i / b_{\bullet} \right\| signo \left\| b_{\bullet} \right\|.$$

- La media total es la proyección del marcador fila medio sobre el vector que representa el vector columna promedio, es decir:

$$x_{\bullet\bullet} \approx a_{\bullet}^{T} b_{\bullet} \approx \left\| proy\, b_{\bullet} / a_{\bullet} \right\| signo \left\| a_{\bullet} \right\| = \left\| proy\, a_{\bullet} / b_{\bullet} \right\| signo \left\| b_{\bullet} \right\|.$$

Propiedades de los marcadores

Las propiedades de los marcadores dependen del tipo de Biplot elegido. Estas propiedades son fundamentales a la hora de elegir el tipo de Biplot que se va utilizar así como la interpretación posterior de los resultados.

Propiedades de los marcadores en el JK-Biplot

En este Biplot se impone la métrica $B^T B = I$, en el espacio de las filas de la matriz X.

- Los productos de los individuos de la matriz X con la métrica identidad, son los productos escalares de los marcadores fila incluidos en A en el espacio completo.

$$XX^T = AA^T.$$

Dado que $X = AB^T$ y $B^T B = I$,

$$XX^T = AB^T BA^T = AA^T.$$

- La distancia Euclídea entre dos individuos de la matriz X es igual a la distancia entre marcadores fila en el espacio completo. Es decir,

$$\left(x_i - x_j\right)^T \left(x_i - x_j\right) = \left(a_i - a_j\right)^T \left(a_i - a_j\right).$$

Como $x_i = Ba_i$,

$$\left(x_i - x_j\right)^T \left(x_i - x_j\right) = \left(Ba_i - \mathrm{B}a_j\right)^T \left(Ba_i - \mathrm{B}a_j\right) =$$

$$\left(a_i - a_j\right)^T B^T B\left(a_i - a_j\right) = \left(a_i - a_j\right)^T \left(a_i - a_j\right).$$

- Los marcadores fila y las coordenadas de los individuos son iguales en el espacio de las componentes principales.

  Si ψ es la matriz que contiene las coordenadas de los individuos en el espacio de las componentes principales entonces $\psi = A$.

$$\psi = XV = \left(UDV^T\right)V = UD = A.$$

- Las coordenadas de las columnas de la matriz X son las proyecciones de los ejes originales sobre el espacio de componentes principales. La proyección de cada marcador fila sobre los marcadores columna es una aproximación de los valores de los individuos sobre las correspondientes variables.
- El producto escalar de marcadores columna es el producto escalar de las columnas de la matriz X con la métrica XX^T. Es decir,

$$x_j^T \left(XX^T\right) x_j = b_j^T b_j .$$

$$x_j^T \left(XX^T\right) x_j =$$

$$b_j^T A^T \left(XX^T\right) A b_j =$$

$$b_j^T b_j .$$

- La similaridad entre columnas se mide utilizando la inversa de la matriz de dispersión de los individuos. Es imposible interpretar los ángulos en términos de correlación debido a:

$$\left(x_i - x_j\right)^T \left(XX^T\right)^{-1} \left(x_i - x_j\right) = \left(b_i - b_j\right)^T \left(b_i - b_j\right).$$

La calidad de representación para las filas es mejor que para las columnas.

Propiedades de los marcadores en el GH-Biplot

En este Biplot se impone la métrica $A^T A = I$.

Los productos escalares de las columnas de la matriz X son iguales a los productos escalares de los marcadores columna.

$$X^T X = BB^T .$$

Dado que $X = AB^T$,

$X^T X = BA^T AB^T = BB^T$.

- Si la matriz X ha sido centrada por columnas, el cuadrado de la longitud de los vectores que representan los marcadores columna aproximan la covarianza

entre las correspondientes variables. Como consecuencia de esta propiedad se derivan las tres propiedades siguientes:

- La longitud al cuadrado del vector que representa un marcador columna aproxima la varianza de la variable correspondiente y la longitud aproxima la desviación estándar.

$$\|b_j\| = \|x_j\| = \sqrt{\text{var}(x_j)}\,.$$

- El coseno del ángulo formado por dos marcadores columna aproxima la correlación entre las variables correspondientes.

$$\cos(b_i, b_j) = corr(x_i, x_j)\,.$$

- La distancia Euclídea entre dos variables es la distancia entre los correspondientes marcadores columna.

$$d^2(x_i, x_j) = (x_i - x_j)^T (x_i - x_j) = \|x_i\|^2 + \|x_j\|^2 - 2(x_i^T x_j) =$$

$$\|b_i\|^2 + \|b_j\|^2 - 2(b_i^T b_j) = d^2(b_i, b_j).$$

- Las coordenadas en la matriz de marcadores columna B son la importancia de las variables en los ejes principales.
- La distancia de Mahalanobis entre dos individuos de la matriz X se aproxima por la distancia Euclídea entre marcadores fila.

$$(x_i - x_j)^T S^{-1} (x_i - x_j) = (a_i - a_j)^T (a_i - a_j)\,.$$

x_i puede ser escrito como Ba_i, entonces:

$$(x_i - x_j)^T S^{-1} (x_i - x_j) =$$

$$(Ba_i - Ba_j)^T S^{-1} (Ba_i - Ba_j) =$$

$$(a_i - a_j)^T B^T (B^T B)^{-1} B (a_i - a_j) = (a_i - a_j)^T (a_i - a_j)\,.$$

Si la matriz X ha sido centrada por columnas, las coordenadas de los marcadores fila son las coordenadas de los individuos en el espacio de las componentes principales. Por lo tanto, la matriz A contiene los scores en las componentes principales estandarizadas.

- Los productos escalares de los marcadores fila son iguales que los productos escalares de las filas de la matriz X con la métrica $\left(X^T X\right)^{-1}$ en el espacio de las columnas.

$$X\left(X^T X\right)^{-1} X^T = AA^T .$$

Dado que $X = AB^T$,

$X\left(X^T X\right)^{-1} X^T = \mathrm{AB}^{\mathrm{T}}\left(B^T B\right)^{-1} BA^T = AA^T$.

La calidad de representación para las columnas es mejor que para las filas.

Propiedades de los marcadores en el HJ-Biplot

(Galindo, 1986) propone un nuevo tipo de Biplot denominado HJ-Biplot que tiene por objetivo una óptima calidad de representación tanto de filas como de columnas aunque el producto interno de sus marcadores no reproduce el elemento de partida. Es decir, $x_{ij} \neq a_i^T b_j$. En este caso los marcadores para las filas de X se calculan como:

$$A = UD$$

y para las columnas como:

$$\mathrm{B} = VD .$$

Las propiedades para los marcadores fila son las descritas en el JK-Biplot y las propiedades para los marcadores columnas son las descritas en el GH-Biplot.

Adicionalmente hay más propiedades acerca de la representación conjunta de filas y columnas.

Marcadores fila y marcadores columna pueden ser representados en el mismo sistema de referencia con óptima calidad de representación. (Greenacre, 1984; Lebart, Morineau, & Piron, 1995) en el contexto del análisis de correspondencias, demuestran que las nubes de los puntos que representan filas y columnas tienen los mismos valores propios y existen relaciones baricéntricas entre ellos. Las relaciones propuestas por (Galindo, 1986) son similares:

Las relaciones entre los vectores propios U y V son:

$U = XVD^{-1}$ y $V = X^T UD^{-1}$. Por lo tanto, los marcadores pueden ser escritos como:

$$A = VD = X^T UD^{-1} D = X^T U = X^T XVD^{-1} = X^T BD^{-1} .$$

$$B = UD = XVD^{-1} D = XV = XX^T UD^{-1} = XAD^{-1} .$$

Por tanto, las coordenadas de las filas son medias ponderadas de las columnas donde los pesos son los valores de la matriz X . Lo mismo aplicado para las coordenadas de las columnas.

Calidad de Aproximación

Para evaluar la calidad de la aproximación S-dimensional es necesario saber qué cantidad de la variabilidad original contenida en la matriz X es explicada por la matriz aproximada X. La variabilidad total de una matriz es la suma de todos sus elementos al cuadrado:

Variabilidad Total = $\|X\|^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^{\ 2}$.

Por las propiedades de la descomposición en valores singulares, esta variabilidad se puede descomponer en una parte explicada y una parte residual:

$$\|X\|^2 = \left\|X\right\|^2 + \left\|X - X\right\|^2 .$$

Usando la ortonormalidad de U y V, podemos expresar esta ecuación en términos de los cuadrados de los valores singulares:

$$\sum_{s=1}^{S} \lambda_s^{\ 2} = \sum_{s=1}^{S} \lambda_s^{\ 2} + \sum_{s=S+1}^{S} \lambda_s^{\ 2} .$$

Esta ecuación muestra que la suma de los S primeros vectores singulares al cuadrado dividido por la suma total de los cuadrados de los valores singulares proporciona una forma de evaluar la cantidad de variabilidad total explicada por los S primeros vectores. Si la cantidad es grande, indica que el gráfico construido por los S primeros vectores singulares da una buena representación de la estructura de la matriz de partida. Si solo una pequeña parte de la variabilidad es explicada por los primeros vectores singulares hay que tener en cuenta que parte de la estructura de la matriz puede estar representada en dimensiones superiores.

Si los datos están centrados por variables, individuos situados cerca del origen de coordenadas pueden tener valores próximos a las medias de las variables o su variabilidad es explicada en otras dimensiones. Del mismo modo, las variables situadas cerca del origen pueden tener una variabilidad pequeña o pueden no estar bien representadas en esas dimensiones.

Para calcular la calidad de representación para columnas se parte de la matriz de varianzas-covarianzas.

$$S = VDU^T UDV^T .$$

En el caso de la representación GH-Biplot, la métrica utilizada es: $A^T A = U^T U = I$. Por lo tanto, la matriz de covarianzas-covarianzas viene dada por:

$$S = VD^2V^T .$$

Es decir, la suma de los cuadrados de los elementos de S es $\sum_{s=1}^{S}\lambda^4$. Por lo que la calidad de representación para las columnas de la aproximación S-dimensional de la matriz X se calcula como:

$$\frac{\sum_{s=1}^{S}\lambda^4}{\sum_{s=1}^{S}\lambda^4}.$$

La calidad de representación para las filas es: S/S ya que la suma de cuadrados de los elementos de $XS^{-1}X^T$ sobre el espacio de las filas de X es igual a S y el de su aproximación es S. En el caso de la representación JK-Biplot, la calidad de representación para columnas es: S/S y la calidad de representación para las filas es:

$$\frac{\sum_{s=1}^{S}\lambda^4}{\sum_{s=1}^{S}\lambda^4}.$$

En el caso de la representación HJ-Biplot, la calidad de representación tanto para filas como para columnas es óptima e igual a:

$$\frac{\sum_{s=1}^{S}\lambda^4}{\sum_{s=1}^{S}\lambda^4}.$$

En la Tabla 1 se resume la obtención de los marcadores fila y columna para cada tipo de Biplot así como la calidad de representación en cada caso.

	Filas		Columnas	
	Coordenadas	Calidad	Coordenadas	Calidad
GH-Biplot	U	$\frac{S}{S}$	VD	$\frac{\sum_{s=1}^{S}\lambda^4}{\sum_{s=1}^{S}\lambda^4}$
JK-Biplot	UD	$\frac{\sum_{s=1}^{S}\lambda^4}{\sum_{s=1}^{S}\lambda^4}$	V	$\frac{S}{S}$
HJ-Biplot	UD	$\frac{\sum_{s=1}^{S}\lambda^4}{\sum_{s=1}^{S}\lambda^4}$	VD	$\frac{\sum_{s=1}^{S}\lambda^4}{\sum_{s=1}^{S}\lambda^4}$

Tabla 1. Obtención de marcadores y sus calidades de representación

Contribuciones

Las calidades de representación explicadas anteriormente son una forma de evaluar globalmente los ajustes de la aproximación pero también es posible medir el ajuste de

individuos y variables a nivel individual, lo cual es importante a la hora de interpretar los resultados. Estas medidas individuales están basadas en los conceptos de contribución (absoluta y relativa) (Benzécri, 1973; Galindo & Cuadras, 1986; Galindo, 1986; Jambu, 1991).

Se tiene:

$$\text{Varianza total de la nube de individuos} =$$

$$\text{Varianza total de la nube de variables} =$$

$$traza\left(XX^T\right) = traza\left(X^T X\right) =$$

$$\sum_{s=1}^{S} \lambda_s{}^2 =$$

$$\sum_{j=1}^{J} d^2\left(b_j, 0\right) = \sum_{s=1}^{S} \sum_{j=1}^{J} b_{js}{}^2 =$$

$$\sum_{i=1}^{I} d^2\left(a_i, 0\right) = \sum_{s=1}^{S} \sum_{i=1}^{I} a_{is}{}^2 .$$

Las contribuciones absolutas de los individuos para la varianza del eje s:

$$CAE_i F_s = a_{is}{}^2 .$$

Las contribuciones absolutas de las variables para la varianza del eje s:

$$CAE_j F_s = b_{js}{}^2 .$$

La inercia total del factor s considerando las contribuciones de los individuos:

$$\sum_{i=1}^{I} a_{is}{}^2 = \lambda_s{}^2 .$$

La inercia total del factor s considerando las contribuciones de las variables:

$$\sum_{j=1}^{J} b_{js}{}^2 = \lambda_s{}^2 .$$

La contribución relativa del elemento i al factor s:

$$CRE_i F_s = \frac{CAE_i F_s}{\lambda_s} .$$

La contribución relativa del elemento j al factor s:

$$CRE_j F_s = \frac{CAE_j F_s}{\lambda_s} .$$

La contribución relativa del factor s al elemento i:

$$CRF_s E_i = \frac{a_{is}{}^2}{d^2\left(a_i, 0\right)} = \cos^2\left(\alpha_i\right).$$

La contribución relativa del factor s al elemento j:

$$CRF_s E_j = \frac{b_{js}{}^2}{d^2\left(b_j, 0\right)} = \cos^2\left(\beta_j\right).$$

Las contribuciones relativas del elemento al factor evalúan en qué medida ese factor puede ser explicado por ese individuo o esa variable.
Las contribuciones relativas del factor al elemento evalúan en qué medida el significado de ese factor puede estar relacionado con el significado de ese individuo o esa variable.

Metodología Bootstrap

Introducción

La teoría estadística intenta responder a tres preguntas básicas:

1. ¿Cómo debo recoger mis datos?
2. ¿Cómo debo recoger y analizar los datos recogidos?
3. ¿Con qué precisión se han analizado y resumido los datos recogidos?

La tercera cuestión corresponde a lo que se conoce como inferencia estadística. Los métodos bootstrap (Efron & Tibshirani, 1993; Efron, 1979, 1987) han sido desarrollados recientemente para hacer algunos tipos de inferencia. Esta técnica requiere del uso de ordenadores para simplificar los cálculos que lleva a cabo.

El uso del término bootstrap proviene del dicho *to pull oneself up by one´s bootstrap,* basado en las aventuras del Barón Munchausen de Rudolph Eric Raspe.

La técnica bootstrap se basa en la elección de sucesivas muestras aleatorias de tamaño n. Dada una población con N unidades, se define una muestra aleatoria simple de tamaño n al conjunto de n unidades escogidas de la población de partida en la que cada unidad del 1 a la N tiene una probabilidad de ser escogida de $1/N$. Este tipo de muestreo se realiza con reposición, es decir, cada unidad de la población puede ser elegida más de una vez.

El principio plug-in

El principio plug-in es una manera sencilla de estimar parámetros a partir de una muestra. Si $\theta = t(F)$, entonces el estimador plug-in es:

$$\hat{\theta} = t(\hat{F})$$

Es decir, se estima el parámetro a calcular a partir de la distribución empírica en lugar de la distribución de probabilidad.

Los métodos bootstrap se utilizan para estudiar el sesgo y el error estándar de estimadores plug-in con la ventaja de calcular el sesgo y el error estándar de una manera automática sin tener en cuenta lo complicado que pueda resultar el cálculo del estimador.

Errores estándar y errores estándar estimados

Supongamos que tenemos una variable aleatoria con distribución de probabilidad F y cuya media y varianza poblacionales son μ_F y σ_F^2 respectivamente.

Sea $x_1, x_2, ..., x_n$ una muestra aleatoria simple de tamaño n de dicha distribución. La media de la muestra $\overline{x}$ tiene una esperanza μ_F y una varianza σ_F^2 / n.

El error estándar de la media $\overline{x}$ es la raíz cuadrada de la varianza de $\overline{x}$.

$$SE_F(\overline{x}) = \sigma_F / \sqrt{n}$$

El error estándar es la manera más usual de explicar precisión estadística. Según el teorema central del límite, bajo ciertas condiciones generales de F, la distribución de $\overline{x}$ se aproxima a una normal cuando n es suficientemente grande, es decir:

$$\overline{x} \sim N(\mu_F, \sigma_F^2 / n)$$

De aquí podemos deducir que la probabilidad de que la diferencia entre la media muestral y la poblacional sea menor que un error estándar es del 68.3%; de igual modo la probabilidad de que dicha diferencia sea menor que 2 errores estándar es del 95.4%.

$$\text{Prob}(|\overline{x} - \mu_F| < \sigma_F / \sqrt{n}) \approx 0.683$$

$$\text{Prob}(|\overline{x} - \mu_F| < 2\sigma_F / \sqrt{n}) \approx 0.954$$

Una de las ventajas de los métodos bootstrap es la capacidad de calcular estas probabilidades a partir de los datos sin basarse en el teorema central del límite con sus respectivas condiciones sobre la función de distribución.

El estimador bootstrap del error estándar

Supongamos que nos encontramos en la siguiente situación: Se ha observado una muestra aleatoria $x=(x_1,x_2,...,x_n)$ proveniente de una población cuya distribución de probabilidad F es desconocida y se quiere estimar un parámetro de interés a partir de la muestra x. La siguiente cuestión que se plantea es cómo de preciso es dicho estimador. Los métodos bootstrap fueron creados en 1979 (Efron, 1979) con el objetivo de calcular el error estándar de un estimador. Es un método completamente automático y puede ser utilizado con cualquier estimador independientemente de la complejidad matemática del mismo.

Los métodos bootstrap están basados en la idea de *muestra bootstrap*. Si se parte de una muestra aleatoria $x=(x_1,x_2,...,x_n)$, una muestra bootstrap $x^*=(x^*_1,x^*_2,...,x^*_n)$ es una muestra aleatoria simple con reposición de n unidades procedentes de la muestra de partida.

Si definimos $\hat{\theta}=s(x)$ como el parámetro de interés a estimar, se puede denominar réplica bootstrap a $\hat{\theta}^*=s(x^*)$ que es el cálculo del estimador a partir de la muestra bootstrap.

Para calcular la precisión del estimador $\hat{\theta}$ se calcula el estimador bootstrap del error estándar de $\hat{\theta}$, $se_{\hat{F}}(\hat{\theta}^*)$ que es el error estándar de $\hat{\theta}$ calculado a partir de muestras aleatorias de tamaño n procedentes de x. Esta estimación se suele llamar *estimador bootstrap ideal del error estándar de* $\hat{\theta}$. Para conseguir una buena aproximación al valor numérico de dicho error se utiliza el algoritmo bootstrap definido a continuación:

1. Se seleccionan B muestras bootstrap independientes $x^{*1},x^{*2},...,x^{*B}$ en la que cada una contiene n unidades escogidas con reposición de x.
2. Se evalúa la réplica bootstrap de cada muestra bootstrap,

$$\hat{\theta}^*(b)=s(x^{*b}) \qquad b=1,2,...,B$$

3. Se estima el error estándar $se_F(\hat{\theta})$ como la desviación estándar de las B réplicas.

$$se_B = \left\{ \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \hat{\theta}^*(\bullet) \right]^2 \Big/ B-1 \right\}^{1/2},$$

donde $\hat{\theta}^*(\bullet) = \sum_{b=1}^{B} \hat{\theta}^*(b) \Big/ B$.

Mientras mayor sea B más se aproxima el estimador del error estándar a su valor real. Este tipo de bootstrap se denomina *bootstrap no paramétrico* ya que utiliza un estimador no paramétrico de la población F.

El algoritmo bootstrap explicado anteriormente está basado en la estructura de datos más simple: una muestra procedente de una población con distribución de probabilidad desconocida F. Sin embargo, hay análisis estadísticos que requieren del uso de estructuras más complejas como las series temporales, el análisis de la variaza, modelos de regresión… El algoritmo bootstrap puede ser adaptado para poder utilizarse en el caso de tener unos datos cuya estructura es más compleja.

El estimador del sesgo

Hasta ahora se ha utilizado el error estándar como medida de precisión de un estimador $\hat{\theta}$ pero existen otras medidas que analizan diferentes aspectos del comportamiento de un estimador. Entre ellos podemos mencionar el sesgo, que es la diferencia entre la esperanza de un estimador y el valor que se desea estimar. El algoritmo bootstrap se puede adaptar fácilmente para estimar el sesgo.

Sea de nuevo F una función de distribución desconocida y sea x una muestra aleatoria de tamaño n procedente de F. Se desea estimar el valor real de un parámetro $\theta = t(F)$.

El sesgo del estimador $\hat{\theta} = s(x)$ se define como la diferencia entre la esperanza del estimador y el valor del parámetro estimado,

$$sesgo_F(\hat{\theta}, \theta) = E_F\left[s(x)\right] - t(F)$$

Los estimadores insesgados juegan un papel importante en la teoría y la práctica estadística. Los estimadores plug-in suelen tener pequeños sesgos en comparación con sus errores estándar lo cual es una buena característica del principio plug-in.

El estimador bootstrap del sesgo se define como:

$$sesgo_{\hat{F}}(\hat{\theta},\theta) = E_{\hat{F}}\left[s(x^*)\right] - t(\hat{F})$$

Para calcular el estimador bootstrap se generan B muestras bootstrap, se calcula para cada una de ellas el estimador $\hat{\theta}^{*b} = s(x^{*b})$ y se aproxima la esperanza $E_{\hat{F}}\left[s(x^*)\right]$ mediante la media:

$$\hat{\theta}^*(\bullet) = \sum_{b=1}^{B} \hat{\theta}^*(b) \Big/ B = \sum_{b=1}^{B} s(x^{*b}) \Big/ B$$

El estimador bootstrap del sesgo sería:

$$\text{sesgo}_{\text{B}} = \hat{\theta}^*(\bullet) - \text{t}(\hat{\text{F}}) .$$

El Jackknife

El jackknife es un método original para estimar sesgos y errores estándar (Quenouille, 1949; Tukey, 1958). Supongamos que disponemos de una muestra $x = (x_1, x_2, ..., x_n)$ y un estimador $\hat{\theta} = s(x)$ y se desea estimar el sesgo y el error estándar del estimador. Se denomina i-ésima muestra jackknife $x_{(i)}$ a la muestra resultante de eliminar de la muestra de partida la i-ésima observación. El estimador jackknife del sesgo se defino como:

$$sesgo_{jack} = (n-1)(\hat{\theta}_{(\bullet)} - \hat{\theta})$$

Donde $\hat{\theta}_{(\bullet)} = \sum_{i=1}^{n} \hat{\theta}_{(i)} \big/ n$ y $\hat{\theta}_{(i)} = s(x_{(i)})$.

El estimador jackknife del estándar error se define como:

$$se_{jack} = \left[\frac{n-1}{n} \sum (\hat{\theta}_{(i)} - \hat{\theta}_{(\bullet)})^2 \right]^{1/2} .$$

La pregunta a realizar ahora es: ¿cuál es mejor, bootstrap o jackknife? Desde el punto de vista computacional, el método jackknife es mejor para tamaños de muestra pequeños porque sólo necesita n muestras frente a las 100 o 200 que suele utilizar el

bootstrap. En el caso del estimador jackknife del error estándar la precisión depende de cómo de cerca esté el estimador de ser lineal. En el caso de tratarse del estimador jackknife del sesgo se pueden aceptar resultados similares para ambos métodos en relaciones cuadráticas.

Intervalos de confianza basados en resultados bootstrap

Hasta ahora nos hemos centrado en el cálculo de errores estándar mediante métodos bootstrap, pero estos errores se utilizan frecuentemente para calcular intervalos de confianza aproximados para un parámetro de interés.

Supongamos que tenemos un estimador $\hat{\theta}$ que sigue una distribución normal de media desconocida θ y varianza conocida se^2,

$$\hat{\theta} \sim N(\theta, se^2).$$

Entonces $\hat{\theta} - \theta \big/ se$ sigue una distribución normal estándar,

$$Z = \frac{\hat{\theta} - \theta}{se} \sim N(0,1).$$

La ecuación $\text{Prob}\left\{|Z| \leq z^{(1-\alpha)}\right\} = 1 - 2\alpha$ es equivalente a $\text{Prob}_{\theta}\left\{\theta \in \left[\hat{\theta} - z^{(1-\alpha)}se, \hat{\theta} - z^{(\alpha)}se\right]\right\} = 1 - 2\alpha$.

Si denotamos $\hat{\theta} - z^{(1-\alpha)}se$ como $\hat{\theta}_{\text{inf}}$ y $\hat{\theta} - z^{(\alpha)}se$ como $\hat{\theta}_{\text{sup}}$, el intervalo $\left[\hat{\theta}_{\text{inf}}, \hat{\theta}_{\text{sup}}\right]$ tiene exactamente una probabilidad de $1 - 2\alpha$ de contener el verdadero valor de θ.

Estos resultados son válidos para un tamaño de muestra suficientemente grande ($n \geq 20$). Para muestras de tamaño menor se tiene:

$$Z = \frac{\hat{\theta} - \theta}{se} \sim \text{t}_{\text{n}-1}.$$

Es decir, utilizando esta aproximación el intervalo de confianza para muestra pequeñas sería:

$$\left[\hat{\theta} - \text{t}_{n-1}^{(1-\alpha)}se, \hat{\theta} - \text{t}_{n-1}^{(\alpha)}se\right]$$

Donde $t_{n-1}^{(\alpha)}$ es el α-ésimo percentil de la distribución t de Student con $n-1$ grados de libertad.

El intervalo t-bootstrap

La distribución t de Student no ajusta los intervalos de confianza si se quiere tener en cuenta la asimetría y otros aspectos en el caso de que el estimador no sea la media muestral. El intervalo t-bootstrap es un procedimiento que ajusta estos errores.

Los métodos bootstrap obtienen intervalos de confianza precisos sin tener en cuenta la suposición de normalidad de los datos. La aproximación t-bootstrap estima la distribución directamente de los datos y construye una tabla de percentiles para posteriormente calcular los intervalos de confianza de la misma forma que en el caso de la distribución normal o la t de Student.

El algoritmo es el siguiente:

1. Se seleccionan B muestras bootstrap independientes $x^{*1}, x^{*2}, ..., x^{*B}$ en la que cada una contiene n unidades escogidas con reposición de x.

2. Se evalúa la réplica bootstrap de cada muestra bootstrap,

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{se^*(b)}, \qquad b = 1, 2, ..., B$$

Donde $\hat{\theta}^*(b) = s(x^{*b})$ es el valor de $\hat{\theta}$ para la muestra bootstrap x^{*b} y $se^*(b)$ es el error estándar estimado de $\hat{\theta}^*$ para dicha muestra.

3. Se estima el α-ésimo percentil de $Z^*(b)$ mediante el valor $\hat{t}^{(\alpha)}$ que satisface:

$$\#\left\{Z^*(b) \le \hat{t}^{(\alpha)}\right\} \Big/ B = \alpha$$

4. Finalmente, se calcula el intervalo de confianza como:

$$\left[\hat{\theta} - \hat{t}^{(1-\alpha)} se, \hat{\theta} - \hat{t}^{(\alpha)} se\right].$$

Si $B\alpha$ no es entero, se puede asumir $\alpha \le 0.5$ y calcular k como el mayor entero que cumple $k \le (B+1)\alpha$. A partir de aquí se calculan α y $1-\alpha$ como el k-ésimo y el $B+1-k$-ésimo mayores valores de $Z^*(b)$ respectivamente.

Intervalos de confianza basados en percentiles bootstrap

Sea $\hat{\theta}$ el estimador plug-in del parámetro θ y se el error estándar estimado. Consideramos el intervalo de confianza normal estándar $\left[\hat{\theta} - z^{(1-\alpha)} se, \hat{\theta} - z^{(\alpha)} se\right]$. Si θ^* es una variable aleatoria que sigue una distribución normal $\hat{\theta}^* \sim N(\theta, se^2)$ podemos expresar los extremos del intervalo de confianza como $\hat{\theta}_{\text{inf}} = \hat{\theta} - z^{(1-\alpha)} se$ y $\hat{\theta}_{\text{sup}} = \hat{\theta} - z^{(\alpha)} se$, siendo dichos extremos el 100α-ésimo y el $100(1-\alpha)$-ésimo percentiles de θ^*.

Los métodos basados en t-bootstrap y en percentiles obtienen buenos resultados en cuanto a probabilidades teóricas se refiere. Sin embargo, en la práctica son imprevisibles.

Para mejorar estos métodos se desarrolló una versión del método basado en percentiles que se denomina método BCa y cuyas siglas corresponden a *bias-corrected and accelerated.*

Supongamos que $\theta^{*(\alpha)}$ indica el 100α-ésimo percentil de B réplicas bootstrap $\theta^*(1), \theta^*(2), ..., \theta^*(B)$. El intervalo BCa $\left[\hat{\theta}_{\text{inf}}, \hat{\theta}_{\text{sup}}\right]$ que tiene una cobertura de $1-2\alpha$ se obtiene de:

$$\left(\hat{\theta}_{\text{inf}}, \hat{\theta}_{\text{sup}}\right) = \left(\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}\right),$$

Donde:

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a\left(z_0 + z^{(\alpha)}\right)}\right)$$

$$\alpha_2 = \Phi\left(z_0 + \frac{z_0 + z^{(1-\alpha)}}{1 - a\left(z_0 + z^{(1-\alpha)}\right)}\right)$$

Aquí $\Phi(\bullet)$ es la función de distribución normal estándar acumulada y $z^{(\alpha)}$ el 100α-ésimo percentil de la distribución normal estándar. Se puede ver que el método

basado en percentiles es un caso particular del método BCa en el que a y z_0 son cero.

Los números a y z_0 se denominan aceleración y corrección del sesgo. z_0 se obtiene de la proporción de réplicas bootstrap que son menores que el estimador original $\hat{\theta}$:

$$z_0 = \Phi^{-1}\left(\frac{\#\left\{\theta^*(b) < B\right\}}{B} \right).$$

$\Phi^{-1}(\bullet)$ indica la inversa de la función de distribución normal estándar acumulada.

Este parámetro mide la discrepancia entre las medianas de θ^* y $\hat{\theta}$.

Para calcular a hay varias formas, la más fácil de explicar está basada en valores jackknife del estadístico $\hat{\theta}$.

Si $x_{(i)}$ es la muestra original sin el punto i-ésimo, sea $\hat{\theta}_{(i)} = s\left(x_{(i)}\right)$ y $\hat{\theta}_{(\bullet)} = \sum_{i=1}^{n} \hat{\theta}_{(i)} \big/ n$, entonces la aceleración puede expresarse como:

$$a = \frac{\sum_{i=1}^{n} \left(\theta_{(\bullet)} - \theta_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n} \left(\theta_{(\bullet)} - \theta_{(i)}\right)^2\right\}^{3/2}}.$$

Este parámetro se refiere al cambio del error estándar de $\hat{\theta}$ respecto del verdadero valor del parámetro θ.

La principal desventaja de este método es la necesidad de un gran número de réplicas bootstrap. Para mejorar este punto se utiliza el método ABC que proviene de las siglas *approximate bootstrap confidence* intervalos.

Evaluación del error en las estimaciones bootstrap

Hasta ahora se han utilizado los métodos bootstrap para evaluar precisión estadística. Los estimadores bootstrap son muy próximos a ser insesgados debido a como se han construido pero tienen una cierta varianza. Esta varianza proviene de dos fuentes diferentes: por un lado se tiene la variabilidad muestral, debida a que sólo se dispone

de una muestra de tamaño n en lugar de la población entera; por otro lado se tiene la variabilidad del remuestreo bootstrap, debida a que sólo se toman B muestras en lugar de infinitas.

Nos centraremos en la estimación bootstrap del error estándar para $s(x)$ donde el error estándar se definía como:

$$se_B = \left\{ \sum_{b=1}^{B} \left[s(x^{*b}) - \overline{s} \right]^2 \Big/ B \right\}^{1/2},$$

donde $\overline{s} = \sum_{b=1}^{B} s(x^{*b}) \Big/ B$.

Esta cantidad tiene una variabilidad que se puede aproximar mediante:

$$\operatorname{var}\left(se_B \right) \doteq \frac{c_1}{n^2} + \frac{c_2}{nB},$$

Donde c_1 y c_2 son constantes que dependen de la población de la cual proviene la muestra pero no dependen de n ni de B. El primer sumando se refiere a la variabilidad de la muestra y tiende a cero cuando el tamaño de la muestra se aproxima a infinito; el segundo sumando cuantifica la variabilidad del remuestreo bootstrap y tiende a cero cuando B tiende a infinito con n fijo. Esta variabilidad nos ayuda para determinar el número de réplicas bootstrap necesario.

Para ello se considera el coeficiente de variación del error estándar:

$$cv\left(se_B \right) = \frac{\operatorname{var}\left(se_B \right)^{1/2}}{E\left(se_B \right)}.$$

Esta expresión se puede escribir como:

$$cv\left(se_B \right) = \left\{ cv\left(se_\infty \right)^2 + \frac{E\left(\hat{\Delta} \right) + 2}{4B} \right\}^{1/2}$$

Donde $\hat{\Delta}$ es la curtosis de la distribución de $\hat{\theta}$ y se_∞ es el estimador bootstrap ideal del error estándar.

Supongamos que tenemos B muestras bootstrap y que hemos calculado el estimador el error estándar de $s(x)$. Se desea medir la incertidumbre en se_B. El método *jackknife-after-bootstrap* proporciona una manera de estimar la variabilidad del error estándar a partir de la información de las B muestras bootstrap únicamente. El método se puede resumir en los siguientes pasos:

1. Para $i = 1, 2, ..., n$ calcular se_B sin utilizar el punto i y se denomina $se_{B(i)}$.

2. Definir $\text{var}_{jack}\left(se_B\right) = \left[n-1/n\right]\sum_{i=1}^{n}\left(se_{B(i)} - se_{B(\bullet)}\right)^2$

   Donde $se_{B(\bullet)} = \sum_{i=1}^{n} se_{B(i)} \Big/ n$.

En la práctica, para calcular $se_{B(i)}$ se utilizan aquellas muestras bootstrap que no contengan al punto i.

Lenguaje R

Tanto para obtener la matriz de datos simulados como para realizar los cálculos con los que se obtienen los resultados de la combinación de los métodos bootstrap con el análisis HJ-Biplot se usó el lenguaje de programación R (R Development Core Team, 2012). R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos.

Entre otras características dispone de:

- Almacenamiento y manipulación efectiva de datos.
- Operadores para cálculo sobre variables indexadas (Arrays), en particular matrices.
- Una amplia, coherente e integrada colección de herramientas para análisis de datos.
- Posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora.
- Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas.

El término “entorno” lo caracteriza como un sistema completamente diseñado y coherente, antes que como una agregación incremental de herramientas muy específicas e inflexibles, como ocurre frecuentemente con otros programas de análisis de datos.

Paquetes existentes en R que realizan biplots

La siguiente tabla (Tabla 2) resume los principales paquetes en R que realizan descomposiciones biplot y/o representaciones biplot. En ella se explica brevemente las principales funcionalidades, su fecha de creación y de última modificación y las principales referencias de cada uno. También se ha considerado de interés señalar en qué referencias del desarrollo de las técnicas biplot se han basado para desarrollar cada uno de ellos.

Los métodos biplot descritos anteriormente corresponden al enfoque de (Gabriel, 1971; Galindo, 1986). Sin embargo, (Gower, 1992) propone otros tipos de biplot basándose en la obtención de los marcadores columna a partir de la regresión multivariante. (Gower & Harding, 1988; Gower, 1992) proponen los Biplot no Lineales. (Gower & Hand, 1996) definen los Biplot de Interpolación y Predicción.

Según Gower, los biplots se refieren a la presentación de información de las variables y unidades, de una matriz de datos X. El biplot arquetipo está representado por los ejes de coordenadas cartesianas en el que los ejes calibrados (generalmente ortogonales) representan las variables y las unidades se representan como puntos. Hay dos preguntas que se relacionan con los ejes cartesianos (i) dado un caso de X, que es una fila de X, ¿dónde está el punto P correspondiente y (ii) dado un punto P ¿cuáles son los valores asociados de x?

Los biplots estadísticos pueden diferir de esta definición clásica en varios sentidos. En primer lugar por lo general tienen sólo una aproximación a la representación cartesiana del espacio completo. Esto induce ejes no ortogonales y complica las respuestas a las preguntas (i) y (ii). En segundo lugar, podemos utilizar las métricas basadas en coeficientes de disimilaridad o en variantes de la distancia de Mahalanobis. Por lo tanto, se necesitan extensiones para abarcar metodologías donde X esté representado por diversas formas de escalamiento multidimensional métrico y no métrico. En tercer lugar, es posible que sea necesario incluir variables categóricas en X, tanto nominales como ordinales. Por lo tanto es necesaria una extensión del sistema cartesiano para poder representar las categorías.

El desarrollo de los biplots en el sentido de Gower muestra cómo manejar todas estas situaciones de una manera unificada. La proyección ortogonal es un concepto clave en el uso de ejes cartesianos, el concepto más general del punto más cercano a un conjunto se utiliza para manejar las generalizaciones. Para variables cuantitativas utiliza ejes calibrados, posiblemente no lineales, y para las variables categóricas, conjuntos de puntos marcados que representan a los diferentes niveles de categoría. Los casos siguen siendo representados por puntos.

Paquete	Referencia	Contenido	Fechas
Stats biplot.princomp GABRIEL	Paquete Stats instalado con el módulo base de R	Muestra biplots sobre resultados de un análisis de componentes principales previo (Gabriel)	
calibrate GOWER	Graffelman, J. (2010). calibrate: Calibration of scatterplot and biplot axes. Download from http://cran.r-project.org/web/packages/calibrate/index.html	Este programa calibra vectores de variables en biplots y scatterplots, dibujando marcas a lo largo del vector y etiquetando esas marcas con los valores específicos. La calibración óptima se encuentra utilizando mínimos cuadrados. Una calibración no óptima se puede encontrar si se especifica un factor de calibración.	21-01-2006 20-03-2012
BiplotGUI GOWER	La Grange, A., le Roux, N. and Gardner-Lubbe, S. (2009). BiplotGUI: Interactive biplots in R. Journal of Statistical Software. Download from: http://www.jstatsoft.org/v30/i12/paper Download from http://cran.r-project.org/web/packages/BiplotGUI/index.html	Construye e interactúa con biplots según el libro de Gower Representa las variables como ejes calibrados. Por lo tanto no es posible interpretación de las longitudes de las variables. Mostrar título y cambiarlo Mostrar etiquetas y puntos (con sus valores) o esconderlos. Cambiar el tipo de línea, el color, el tamaño; la fuente, el tamaño, el color y la orientación de las etiquetasy las marcas. Dibujar convex-hulls y alpha bags cuando haya grupos en los datos. Realiza otros tipos de biplot (no lineales, MDS) Permite elegir la distancia a utilizar y la manera de calcular las coordenadas (coordenadas principales, matriz de covarianzas/correlaciones).	13-08-2008 19-03-2013

		Muestra en un gráfico las correlaciones de las variables con los ejes cartesianos. Gráfico en 3D con rotación y zooming	
bpca GABRIEL; GALINDO	Faria, J.C & Demetrio, C. G. B (2012).bpca: Biplot of multivariate data based on Principal Components Analysis.ESALQ, USP, Brasil. Download from http://cran.r-project.org/web/packages/bpca/index.html	• Dibuja **Biplots (Gabriel y Galindo)** en 2D y 3D • Longitudes de variables, ángulos entre variables, correlaciones entre variables. • Coordenadas de individuos y variables, valores y vectores propios. • Calidades de representación de las variables y dibuja las correlaciones y sus aproximadas en el mismo gráfico • Interacción en el gráfico 3D (rotación y zooming)	17-08-2008 21-02-2012
GGEBiplotGUI GABRIEL;GALINDO YANG	E.Frutos Bernal & P.Galindo Villardon (2013). GGEBiplotGUI: Interactive GGE Biplots in R. Salamanca, Spain. Download from http://cran.r-project.org/web/packages/GGEBiplotGUI/index.html Yan W, Hunt LA, Sheng Q, Szlavnics Z (2000). "Cultivar evaluation and mega-environment investigation based on GGE biplot." Crop Sci, 40, 597-605. Yan W, Kang M (2003). "GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists." CRC Press, Boca Raton, FL, USA.	GUI con la que se puede construer e interactuar con **GGE biplots**. Valores propios y porcentaje de variabilidad explicada por cada uno de ellos - Coordenadas de las filas - Coordenadas de las columnas - Contribuciones del factor al elemento (filas y columnas) Se puede: - Cambiar el color de : .Fondo .De todas las etiquetas de los genotipos .De todas las etiquetas de los ambientes .Del título - Se puede cambiar la fuente (opciones default, larger, smaller) - Se puede cambiar el título del gráfico - Se pueden mostrar genotipos, ambientes o ambos. - Mostrar/ ocultar título - Mostrar/ ocultar los ejes - Mostrar/ ocultar símbolos	29-08-2011 22-06-2013

		Además mediante el ratón se pueden mover las etiquetas de genotipos y ambientes (arrastrando) y con el botón derecho del ratón se puede cambiar el color o la etiqueta individualmente para genotipos y ambientes.	
multibiplotGUI GABRIEL; GALINDO	A.B. Nieto Librero & N. Baccala & P. Vicente Galindo & P. Galindo Villardon (2013). multibiplotGUI: Multibiplot Analysis in R. Salamanca, Spain. Download from http://cran.r-project.org/web/packages/multibiplotGUI/index.html	• Biplots (Gabriel y Galindo) • **Multibiplots** • Calidades de representación, contribuciones, bondades de ajuste, valores propios, posibilidad de elegir el número de ejes a retener, gráficos en 2D y 3D, interacción con gráficos (2D: mover etiquetas, cambiarlas o borrarlas, cambiar de color, tamaño y símbolo los puntos representados, elegir las dimensiones a mostrar en el gráfico; 3D: rotar y ampliar gráfico).	29-10-2012

Tabla 2. Software de Biplot en R

También se ha querido mencionar otros paquetes existentes en el entorno R que nombran la palabra biplot, si bien hay que señalar que no utilizan la descomposición biplot sino que se refieren a una representación conjunta de coordenadas resultantes de análisis previos (Tabla 3).

vegan	Download from http://cran.r-project.org/web/packages/vegan/index.html	Proporciona herramientas para describir comunidades ecológicas. Contiene funciones básicas para análisis de diversidad, ordenación y análisis de disimilaridad. Tiene funciones que dibujan biplot a partir de resultados de análisis de Redundancia, análisis de correlación canónica, análisis canónico de correspondencias.	6-09-2001 19-03-2013

ade4	Dray, S. and Dufour, A.B. (2007): The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software. 22(4): 1-20. Chessel, D. and Dufour, A.B. and Thioulouse, J. (2004): The ade4 package-I- One-table methods. R News. 4: 5-10. Dray, S. and Dufour, A.B. and Chessel, D. (2007): The ade4 package-II: Two-table and K-table methods. R News. 7(2): 47-52. Download from http://cran.r-project.org/web/packages/ade4/index.html	ade4 está caracterizado por (1) ls implementación de funciones gráficas y estadísticas, (2) disponible para datos numéricos, (3) la redacción de documentación técnica y (4) la inclusión de referencias bibliográficas. Función que representa biplot de los resultados de los análisis implementados	10-12-2002 11-04-2013
ade4TkGUI	Thioulouse, J. and Dray, S. (2007): Interactive Multivariate Data Analysis in R with the ade4 and ade4TkGUI Packages. Journal of Statistical Software. 22(5): 1-14. Download from http://cran.r-project.org/web/packages/ade4TkGUI/index.html	Una Tcl/Tk GUI para varias funciones básicas del paquete ade4.	29-09-2006 13-11-2012
ca	Nenadic, O., Greenacre, M. (2007) Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. Journal of Statistical Software 20(3):1-13. Download from http://cran.r-project.org/web/packages/ca/index.html	Paquete para la computación y visualización de simple, múltiple y conjunto análisis de correspondencias. Funciones que dibujan biplots a partir de resultados de un análisis de correspondencias	28-07-2007 12-06-2012
caGUI	Download from http://cran.r-project.org/web/packages/caGUI/index.html	Una Tcl/Tk GUI para las funciones del paquete ca.	4-10-2009 29-10-2012
ThreeWay	Download from http://cran.r-	Análisis de Componentes para datos de tres vías (modelos Candecomp/Parafac,	29-10-2012

	project.org/web/packages/ThreeWay/index.html	Tucker3, Tucker2 y Tucker1) Función que dibuja joint biplot de un análisis Tucker3	11-06-2013

Tabla 3. Otros software existentes

Paquete biplotbootGUI

Debido a que se ha considerado que las herramientas existentes no eran adecuadas para nuestro objetivo se ha optado por desarrollar un nuevo paquete que incluye el análisis biplot en el sentido en el que lo desarrollaron Gabriel, 1971; Galindo, 1986 y la metodología bootstrap para presentar sus resultados con medidas de precisión.

El paquete es una interfaz gráfica que permite interactuar mediante el uso de ventanas, botones y menús.

La ventana principal (Figura 5) permite introducir el número de iteraciones que se van a realizar para calcular los intervalos de confianza así como el nivel de confianza para calcularlos. También es posible elegir que se muestren sólo los resultados de las medidas que al usuario le sean de interés.

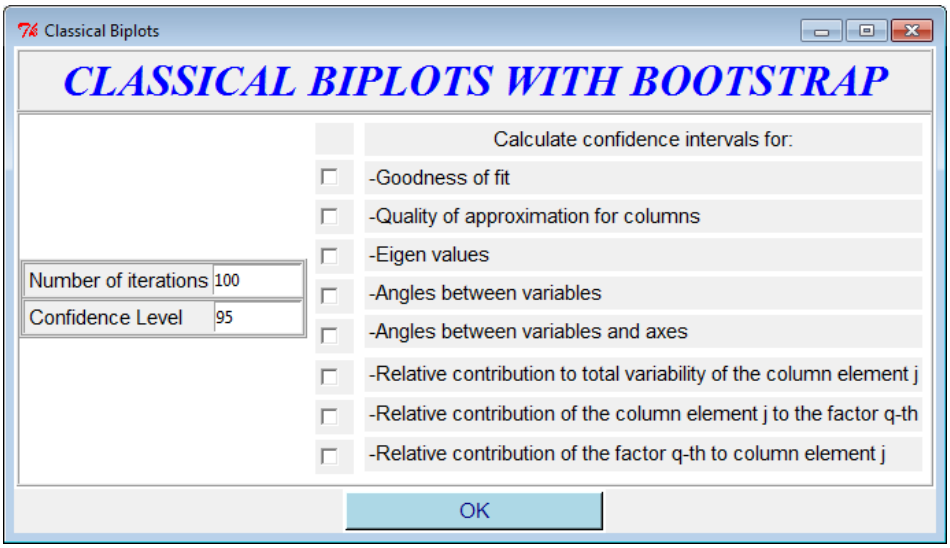


Figura 5. Ventana principal

Una vez que se ha pulsado el botón “OK” aparece la ventana de opciones (Figura 6) en la que se puede:

- Elegir el tipo de biplot que se quiere realizar (HJ, GH o JK).

- Elegir una transformación previa de los datos
    - Restar la media global
    - Centrar por columnas
    - Estandarizar por columnas
    - Centrar por filas
    - Estandarizar por filas
    - No realizar ninguna transformación
- Cambiar color, tamaño, etiqueta y símbolo que van a representar a los individuos en los gráficos.
- Cambiar color, tamaño y etiqueta que van a representar a los variables en los gráficos.
- Mostrar los ejes de coordenadas en los gráficos.

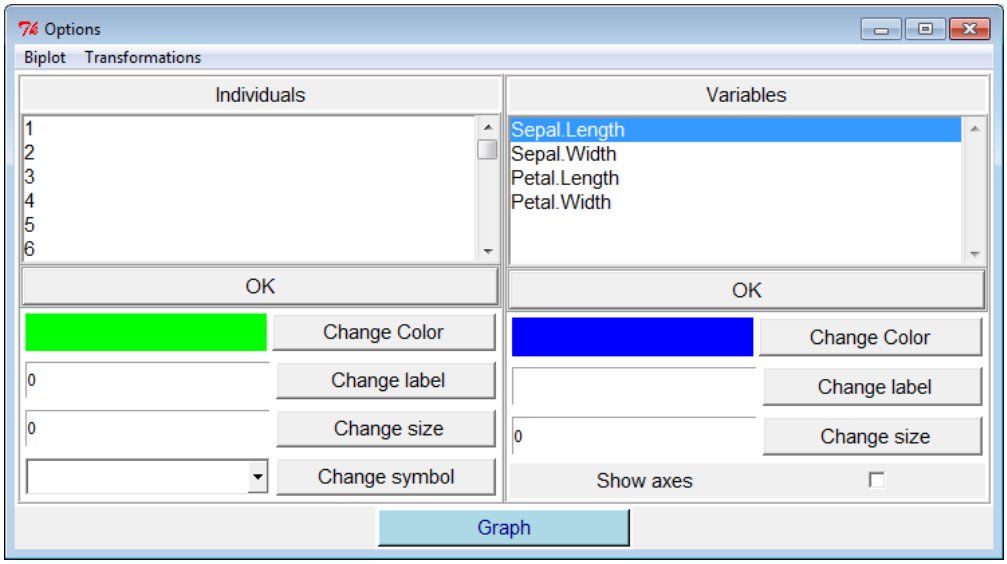


Figura 6. Ventana de Opciones

Una vez que se pulsa el botón “Gráfico” aparece otra ventana (Figura 7) que muestra un diagrama de barras en la que se representa la inercia absorbida por cada eje y se puede elegir el número de ejes a retener para el posterior análisis.

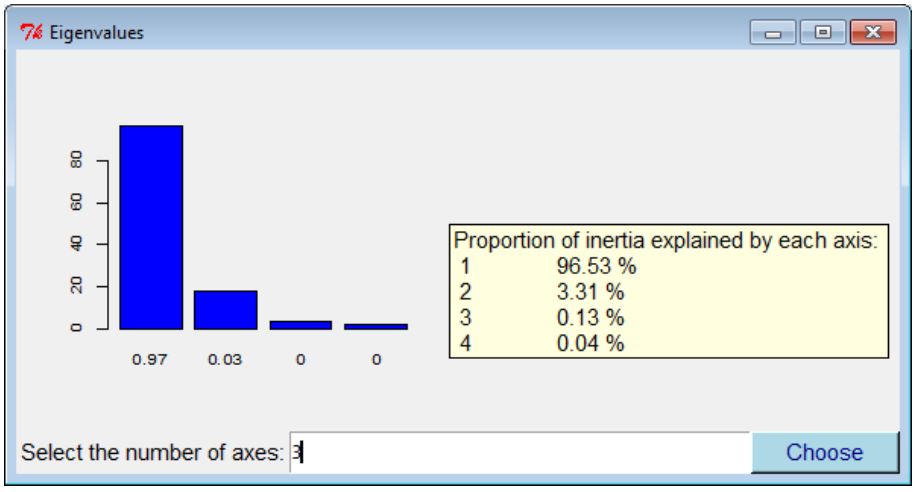


Figura 7. Inercia Absorbida por cada eje

Una vez elegidos el número de ejes y pulsado el botón “Choose” aparece la ventana con el gráfico en dos dimensiones (Figura 8).

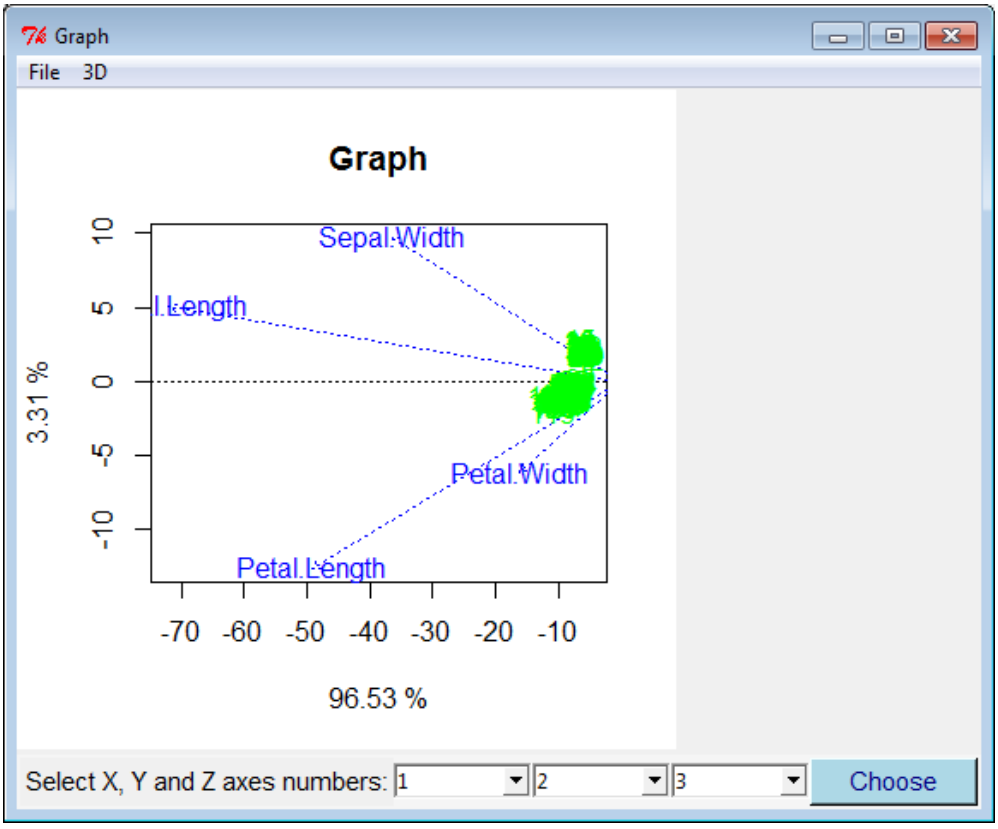


Figura 8. Ventana de Gráfico

En esta ventana es posible elegir los dos ejes que se muestran y un tercero que se puede ver si se elige la opción “3D”. También se pueden mover las etiquetas con el botón izquierdo del ratón y cambiar las características de los puntos con el botón derecho del ratón. Se ha desarrollado también un menú en el que se pueden elegir distintos formatos para guardar el gráfico.

Si nos situamos en el gráfico en tres dimensiones, se puede rotar con el botón izquierdo del ratón y ampliar la imagen con el botón derecho.

Junto al gráfico en dos dimensiones emerge un texto con los resultados del análisis biplot (coordenadas, bondad de ajuste, calidades de representación para filas y columnas, ángulos entre variables y entre variables y ejes, longitudes de los vectores que representan variables, contribuciones de los elementos a la variabilidad total, contribuciones de los elementos a los factores y de los factores a los elementos, inercia absorbida por cada eje y valores propios).

En cuanto a los resultados bootstrap se refiere, aparece otro texto que muestra para cada parámetro elegido el valor observado, la media calculada a partir de los valores del parámetro en cada remuestreo, la desviación estándar, el sesgo y los extremos inferior y superior de los intervalos de confianza t-bootstrap y percentiles.

Estos dos archivos de texto se guardan automáticamente en el directorio en el que se encuentre el usuario así como todos los gráficos que contienen histogramas y gráficos de normalidad de los parámetros seleccionados en la pantalla principal.

Por último, se muestra un gráfico donde aparecen representadas todas las coordenadas de las variables que se han calculado a partir de las muestras bootstrap. Para ello se han realizado rotaciones Procrustes con el objetivo de eliminar el efecto espejo que pudiera existir entre los diferentes conjuntos de coordenadas.

Resultados

Para los datos anteriormente descritos se realizaron 1000 réplicas bootstrap del método HJ-Biplot y se calcularon los intervalos de confianza basados en percentiles con un α=0.5 y los intervalos de confianza t-bootstrap.

Datos Iris

Los resultados para los datos referentes a las flores se detallan a continuación. En primer lugar se realizó un análisis HJ-Biplot con la transformación "Estandarizar por columnas".

En la Tabla 4 se observa la información proporcionada por los valores propios. En ella se aprecia que el primer eje absorbe la mayor parte de la información (más del 70%) y con los tres primeros ejes se explica casi el 100% de la información (99,48%).

Valor Propio	Variabilidad Explicada	Acumulada
20,85	72,96	72,96
11,67	22,85	95,81
4,68	3,67	99,48

Tabla 4. Valores propios y variabilidad explicada datos iris

La siguiente tabla (Tabla 5) recoge las contribuciones relativas del factor al elemento de las diferentes flores que se han analizado en los tres ejes que se han retenido.

	Eje 1	Eje 2	Eje 3
1	954,1	42,86	3,03
2	894,73	93,9	11,37
3	979,18	20,48	0,34
4	935,39	63,14	1,47
5	931,71	68,25	0,04
6	660,1	339,79	0,11
7	981,14	0,37	18,49
8	988,57	9,87	1,56

9	811,63	185,24	3,13
10	943,75	43,51	12,74
11	801,6	186,06	12,33
12	995,13	3,26	1,62
13	893,88	96,44	9,68
14	878,69	117,17	4,14
15	567,56	406,18	26,25
16	414,9	585,03	0,08
17	688,87	311,13	0
18	952,19	47,43	0,39
19	630,3	345,19	24,5
20	809,82	187,59	2,59
21	914,04	41,69	44,27
22	847,05	148,51	4,44
23	960,03	26,2	13,76
24	997,43	2,21	0,36
25	993,44	3,77	2,79
26	887,11	91,14	21,75
27	984,53	13,72	1,75
28	936,17	55,32	8,51
29	963,86	20,66	15,48
30	977,38	21,73	0,89
31	946,25	52,59	1,16
32	929,96	49,76	20,28
33	679,91	319,87	0,22
34	563,64	435,71	0,64

35	948,69	45,13	6,18
36	981,22	8,55	10,23
37	861,5	90,14	48,36
38	947,88	52,06	0,06
39	873,5	120,97	5,53
40	978,54	15,03	6,43
41	963,81	35,97	0,22
42	385,43	609,91	4,66
43	952,8	33,54	13,66
44	923,66	53,4	22,94
45	769,71	219,89	10,4
46	893,67	105,47	0,85
47	818,79	180,74	0,47
48	971,43	25,28	3,3
49	828,5	166,01	5,49
50	995,18	0,02	4,8
51	500,11	306,81	193,08
52	596,08	394,04	9,87
53	691,6	170,57	137,83
54	51,18	948,66	0,16
55	851,29	31,97	116,74
56	291,38	678,87	29,75
57	473,47	507,67	18,85
58	63,65	919,7	16,65
59	707,21	0,85	291,93
60	0,1	786,32	213,58

61	1,72	997,97	0,31
62	808,29	16,67	175,03
63	78,66	775,34	145,99
64	929,3	62,24	8,46
65	4,8	831,3	163,9
66	599,17	202,61	198,22
67	305,04	95,83	599,14
68	33,9	843,46	122,64
69	343,83	602,88	53,29
70	15,51	967,46	17,03
71	503,14	145,41	351,44
72	481,42	369,6	148,98
73	602,05	344,3	53,65
74	607,67	263,06	129,27
75	708,68	5,77	285,55
76	727,55	59,87	212,58
77	747	2,82	250,18
78	912,11	54,26	33,63
79	883,26	102,01	14,73
80	1,32	915,28	83,39
81	6,9	984,03	9,07
82	0,22	976,73	23,05
83	85,13	881,51	33,36
84	731,7	261,09	7,21
85	87,16	143,89	768,95
86	167,08	648,69	184,23

87	719,31	178,24	102,45
88	313,75	550,02	136,24
89	35,08	349,03	615,89
90	43,31	952,38	4,31
91	58,07	935,28	6,66
92	997,34	1,59	1,07
93	100,23	864,59	35,18
94	31,09	966,26	2,65
95	100,02	879,41	20,57
96	144,38	567,88	287,75
97	231,06	660,2	108,74
98	772,61	55,78	171,61
99	76,49	909,67	13,84
100	152,23	828,28	19,5
101	658	146,52	195,48
102	635,34	231,46	133,2
103	931,61	60,51	7,89
104	986,22	1,05	12,73
105	934,74	23,32	41,94
106	885,44	74,91	39,65
107	37,98	687,43	274,59
108	898,15	29,9	71,96
109	858,93	107,96	33,11
110	570,26	412,1	17,65
111	768,4	198,16	33,45
112	935,07	64,74	0,2

113	952,63	47,18	0,19
114	484,79	412,42	102,79
115	641,57	58,26	300,16
116	745,11	134,77	120,12
117	970,09	29,28	0,63
118	473,24	525,45	1,31
119	956,79	0,03	43,18
120	348,59	635,79	15,62
121	824,46	164,59	10,95
122	485,71	166,02	348,27
123	902,58	18,39	79,03
124	884,47	115,51	0,01
125	721,39	256,39	22,21
126	762,14	202,66	35,2
127	921,9	66,83	11,27
128	898,16	3,57	98,28
129	967,23	10,62	22,15
130	807,44	73,5	119,06
131	908,6	10,29	81,11
132	426,59	553,85	19,55
133	956,63	8,79	34,58
134	912,2	63,05	24,74
135	678,43	308,84	12,73
136	883,9	82,84	33,26
137	549,55	252,56	197,88
138	895,65	88,18	16,17

139	830,92	0,29	168,79
140	882,36	117,6	0,04
141	878,71	81,57	39,71
142	880,16	115,72	4,12
143	635,34	231,46	133,2
144	827,68	149,6	22,73
145	726,81	200,38	72,81
146	942	40,32	17,68
147	752,59	247,2	0,21
148	956,65	29,93	13,42
149	498,77	270,65	230,58
150	767,45	0,49	232,06

Tabla 5. Contribuciones relativas del factor al elemento para las flores

Tres de las 150 flores analizadas no están bien representadas en los dos primeros ejes (inferior a 400).

La información acerca de las contribuciones relativas del factor al elemento para las variables se recoge en la Tabla 6.

	Eje 1	Eje 2	Eje 3
Sepal.Length	793,52	130,38	76,09
Sepal.Width	211,8	779,43	8,77
Petal.Length	996,44	0,56	3
Petal.Width	936,5	4,12	59,38

Tabla 6. Contribuciones relativas del factor al elemento para las variables datos iris

Según se puede apreciar, todas las variables están bien representadas en el primer plano principal (plano 1-2).

La representación HJ-Biplot en el primer plano principal se muestra en la Figura 9. En ella se han resaltado en diferentes colores los tres tipos de flores incluidos en la muestra (Rojo-Setosa, Verde-Versicolor, Naranja-Virginica).

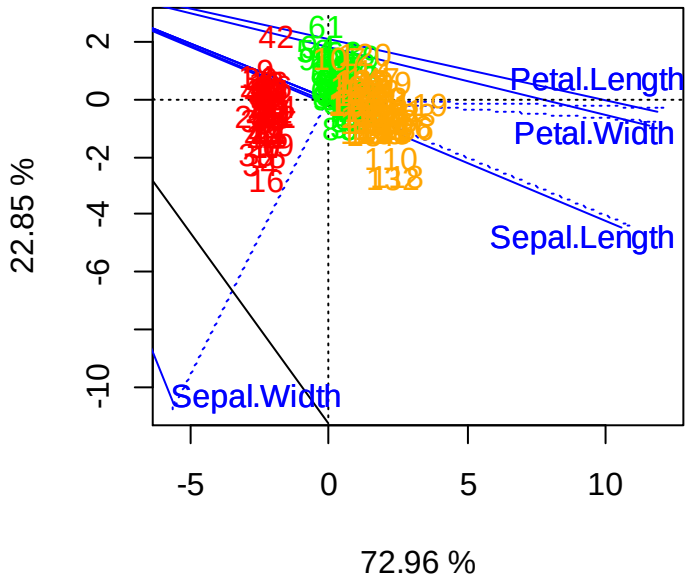


Figura 9. Representación HJ-BIPLOT de los datos iris

La estructura de covariación de las variables pone de manifiesto una alta correlación entre las variables referentes al tamaño y longitud de los pétalos (Petal.Length y Petal.Width) representadas por un ángulo muy pequeño. Ambas variables tienen una correlación alta con la variable Sepal.Length. Sin embargo, no tienen prácticamente relación con la variable Sepal.Width al presentar ángulos próximos a 90º.

Se puede observar que el primer eje separa el grupo setosa del resto caracterizándose por longitudes y tamaños de pétalos más pequeños que los otros dos grupos.

Para analizar con más detalle estas relaciones, se muestran los ángulos entre las variables y entre las variables y los ejes en el plano 1-2 (Tabla 7 y Tabla 8).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0	95,47	20,71	18,27
Sepal.Width	95,47	0	116,18	113,74

Petal.Length	20,71	116,18	0	2,44
Petal.Width	18,27	113,74	2,44	0

Tabla 7. Ángulos entre variables datos iris

	Eje 1	Eje 2
Sepal.Length	22,07	67,93
Sepal.Width	62,47	27,53
Petal.Length	1,35	88,65
Petal.Width	3,79	86,21

Tabla 8. Ángulos entre variables y ejes datos iris

A continuación se muestran los resultados obtenidos de aplicar la metodología bootstrap a los parámetros que se obtienen del HJ-Biplot. Para todos ellos se presentan:

- Histograma de los valores obtenidos de las 1000 réplicas bootstrap. En ellos se ha resaltado en línea continua azul el valor de la media de dichos valores y en línea discontinua roja el valor calculado a partir de la muestra inicial.
- Gráfico de normalidad.
- Tabla en la que se muestra:
    - Valor observado del parámetro (calculado a partir de la muestra inicial).
    - Media de los valores obtenidos a partir de las réplicas bootstrap.
    - Desviación estándar de dichos valores.
    - Sesgo.
    - Extremos inferior y superior del intervalo t-bootstrap.
    - Extremos inferior y superior del intervalo basado en percentiles.

En primer lugar se puede observar los resultados para la calidad de aproximación para las columnas (Figura 10 y Tabla 9). En ella se aprecia que prácticamente no hay diferencia entre el valor observado y el estimado, y que los intervalos de confianza son similares.

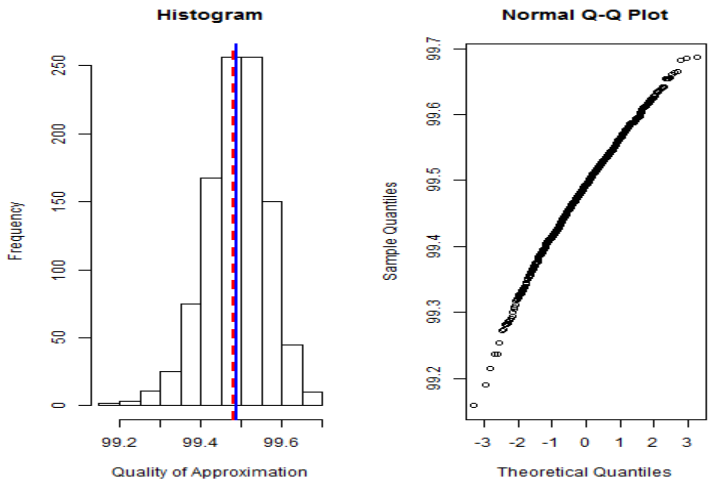


Figura 10. Histograma de la calidad de aproximación datos iris

	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Quality of Approximation	99,48	99,49	0,08	0,01	99,34	99,64	99,33	99,62

Tabla 9. Resultados para calidad de aproximación datos iris

A continuación se muestran los principales resultados para cada uno de los valores propios resultantes de la descomposición (Figura 11 y Tabla 10). En ellos se observan de nuevo que hay una diferencia muy pequeña entre los valores observados y los calculados a través de las réplicas bootstrap y que hay concordancia entre ambos tipos de intervalos.

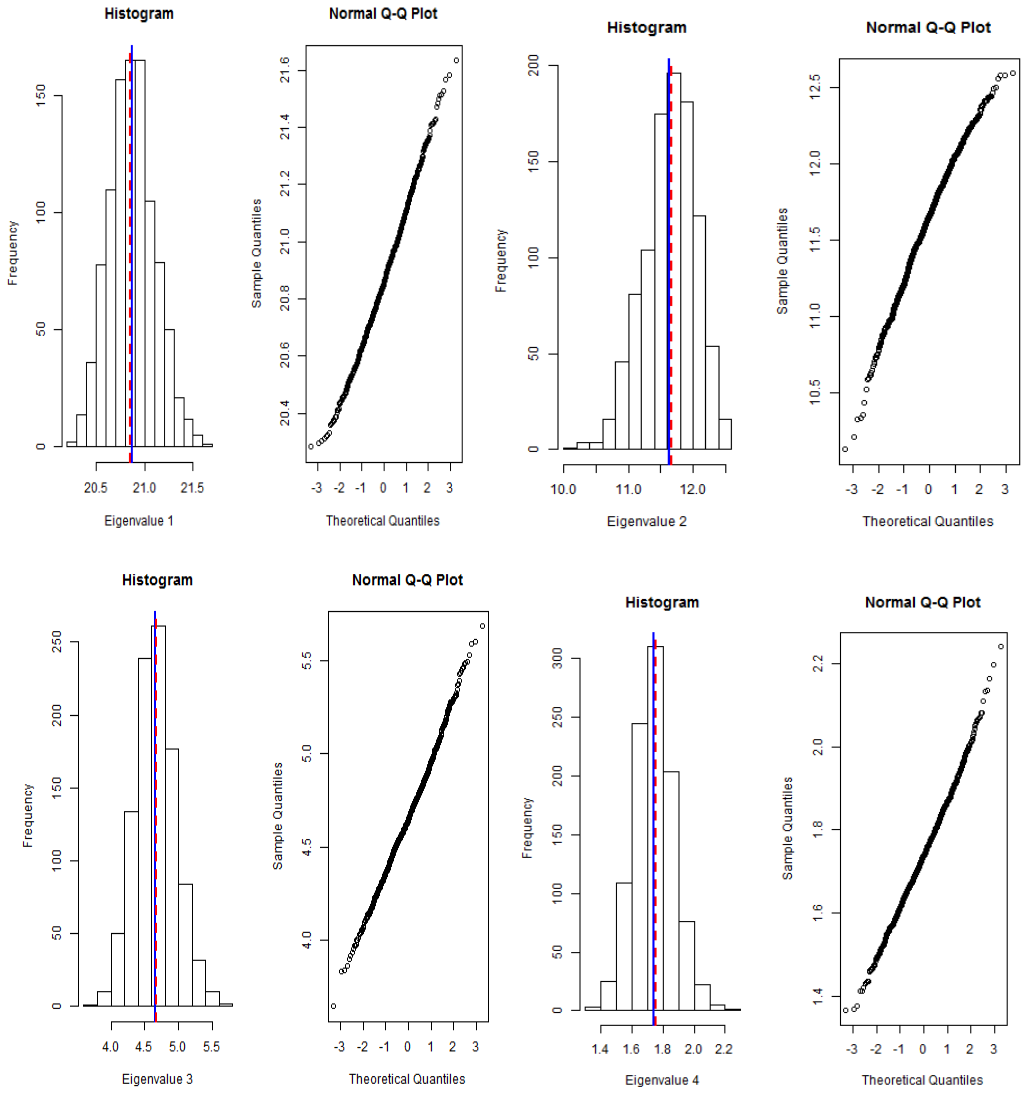


Figura 11. Histograma de los valores propios datos iris

Eigen values:								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Eigenvalue1	20,85	20,87	0,24	0,02	20,4	21,33	20,44	21,35
Eigenvalue2	11,67	11,64	0,4	-0,03	10,84	12,43	10,81	12,32
Eigenvalue3	4,68	4,66	0,3	-0,02	4,06	5,25	4,1	5,28
Eigenvalue4	1,76	1,74	0,13	-0,02	1,49	2	1,5	2

Tabla 10. Resultados para los valores propios datos iris

El siguiente resultado que se muestra es el relativo a los ángulos que forman entre sí las variables (Figura 12 y Tabla 11). Se puede observar, al igual que en los casos anteriores que no existen grandes diferencias entre los valores observados y los calculados y que los intervalos tampoco difieren entre sí. En estos parámetros cabe destacar que existe una ligera asimetría en aquellos ángulos próximos a cero. Esto es debido a que se han realizado transformaciones previas al cálculo de estos resultados con el fin de que todos los ángulos calculados fueran positivos.

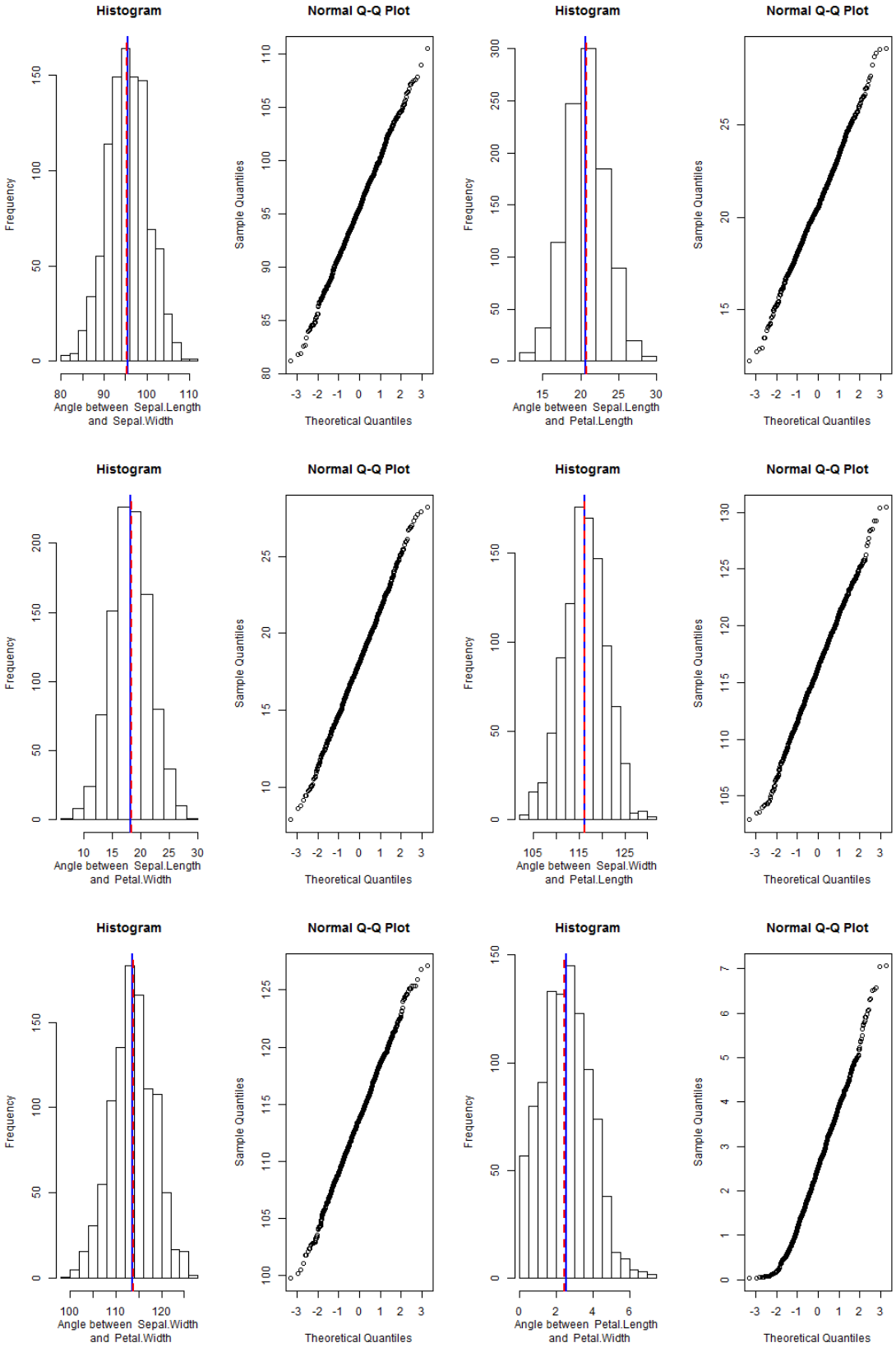


Figura 12. Histograma de ángulos entre variables datos iris

Angles between variables:								
Angles between Sepal.Length and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Sepal.Width	95,47	95,56	4,73	0,09	86,22	104,9	86,52	104,47
Petal.Length	20,71	20,68	2,69	-0,03	15,36	26	15,28	25,93
Petal.Width	18,27	18,16	3,41	-0,12	11,43	24,89	11,45	25,03
Angles between Sepal.Width and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Petal.Length	116,18	116,24	4,64	0,06	107,07	125,41	106,68	124,93
Petal.Width	113,74	113,72	4,68	-0,02	104,47	122,97	104,23	122,53
Angles between Petal.Length and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Petal.Width	2,44	2,55	1,32	0,11	-0,05	5,15	0,21	5,07

Tabla 11. Resultados para ángulos entre variables datos iris

El siguiente parámetro al que se le han calculado intervalos de confianza son los ángulos que forman las variables con los dos primeros ejes (Figura 13 y Tabla 12). Las apreciaciones en este caso son las mismas que para los ángulos entre variables.

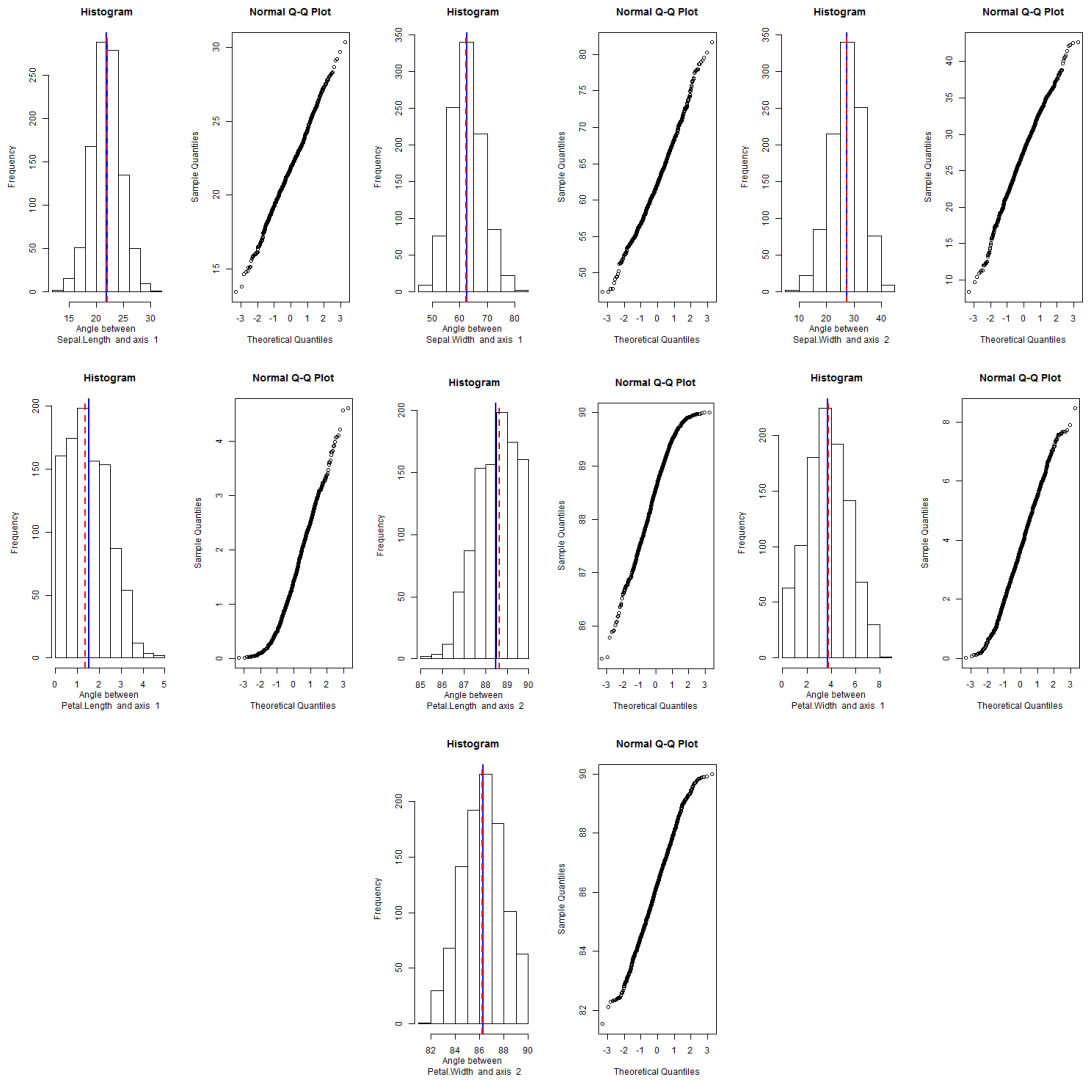


Figura 13. Histograma ángulos entre variables y ejes datos iris

Angles between variables and axes:								
Angles between Sepal.Length and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	22,07	21,83	2,65	-0,24	16,59	27,07	16,48	27,25
Axis 2	67,93	68,17	2,65	0,24	62,93	73,41	62,75	73,52
Angles between Sepal.Width and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup

Axis 1	62,47	62,61	5,72	0,14	51,31	73,9	52,58	74,46
Axis 2	27,53	27,39	5,72	-0,14	16,1	38,69	15,54	37,42
Angles between Petal.Length and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	1,35	1,52	0,93	0,17	-0,31	3,35	0,1	3,37
Axis 2	88,65	88,48	0,93	-0,17	86,65	90,31	86,63	89,9
Angles between Petal.Width and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	3,79	3,74	1,7	-0,06	0,37	7,1	0,65	7,09
Axis 2	86,21	86,26	1,7	0,06	82,9	89,63	82,91	89,35

Tabla 12. Resultados para ángulos entre variables y ejes datos iris

A continuación, se presentan los resultados para la contribución a la variabilidad total de las variables (Figura 14 y Tabla 13). Se observa que entre los valores observados y los calculados no hay prácticamente diferencia y que los intervalos de confianza tienen una amplitud muy pequeña lo que nos sugiere una gran precisión de este parámetro.

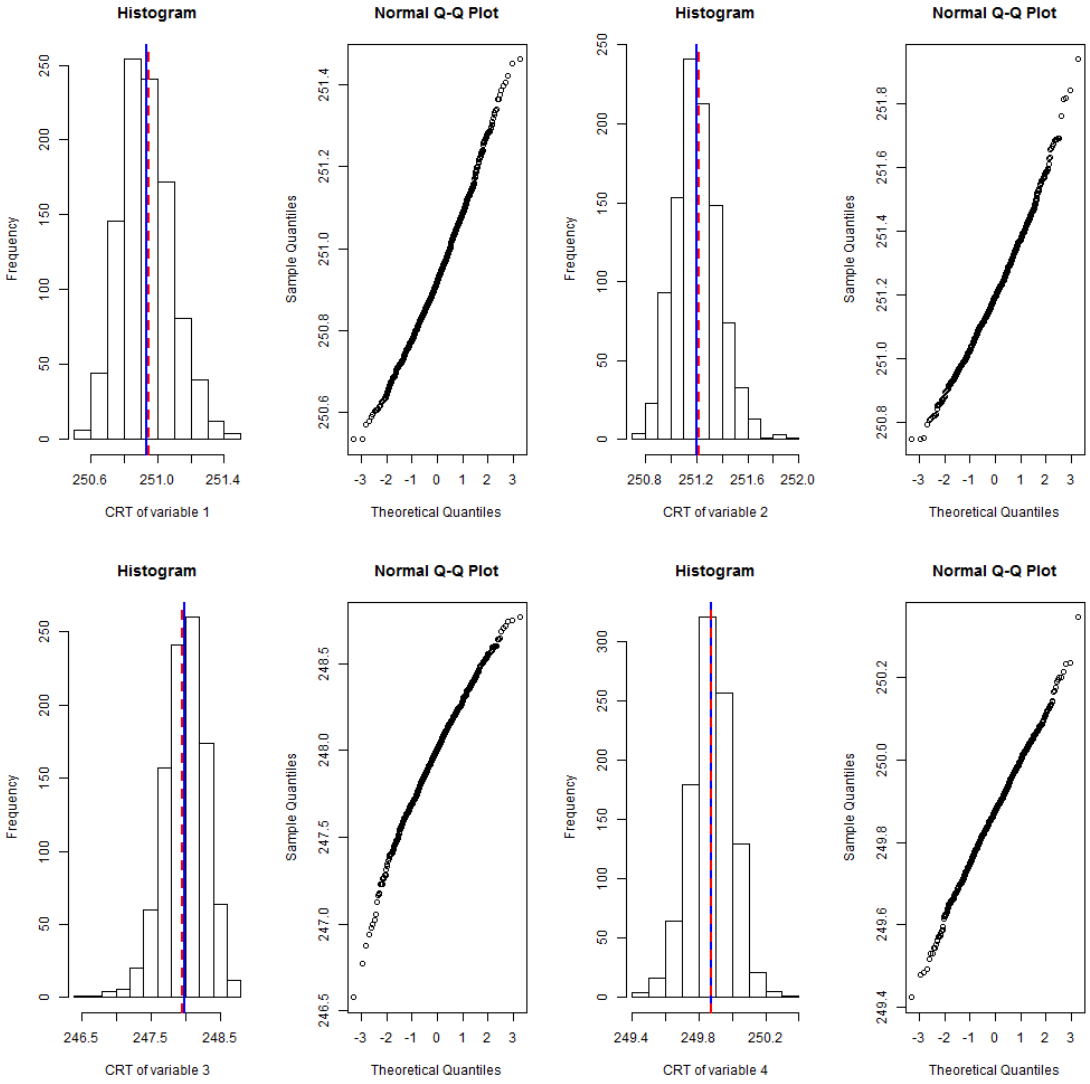


Figura 14. Histograma para contribuciones a la variabilidad total datos iris

Relative contribution to total variability of the column element j:								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Sepal.Length	250,95	250,93	0,16	-0,01	250,62	251,24	250,65	251,27
Sepal.Width	251,22	251,2	0,17	-0,02	250,86	251,55	250,9	251,57
Petal.Length	247,96	247,99	0,31	0,03	247,39	248,59	247,36	248,53
Petal.Width	249,87	249,87	0,13	0	249,63	250,12	249,63	250,1

Tabla 13. Resultados para contribuciones a la variabilidad total datos iris

Los últimos parámetros que se han estimado han sido las contribuciones relativas del factor al elemento y las contribuciones relativas del elemento al factor considerando como elementos a las variables (Figura 15, Figura 16, Tabla 14 y Tabla 15). Nuevamente se puede apreciar que existen pequeñas diferencias entre los valores observados y los calculados a partir del remuestreo bootstrap. En estos parámetros hay que destacar que en algunos casos los intervalos t-bootstrap y los basados en percentiles no concuerdan. Además hay extremos inferiores de los intervalos t-bootstrap que son negativos. Esto es debido a que la distribución de los valores calculados tiene mucha asimetría y los intervalos t-bootstrap no son adecuados en estos casos.

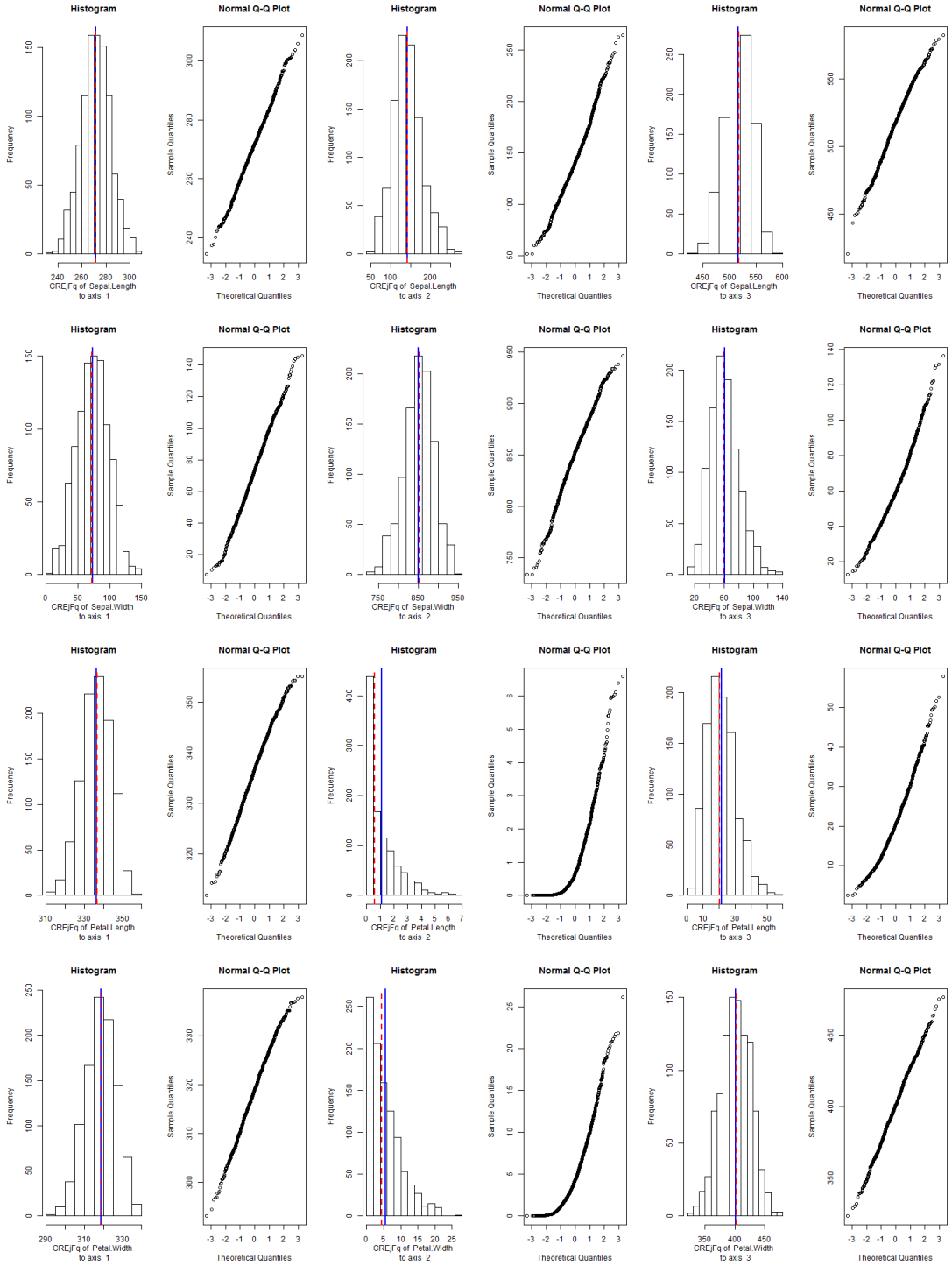


Figura 15. Histograma CREjFq datos iris

Relative contribution of the column element j to the factor q-th:								
Sepal.Length								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	271,51	271,76	12,29	0,25	247,48	296,04	247,34	296,57
Axis 2	142,44	143,15	36,73	0,71	70,57	215,73	74,61	223,48
Axis 3	517,78	516,7	25,58	-1,07	466,16	567,25	467,19	560,29
Sepal.Width								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	72,55	73,19	25,05	0,64	23,68	122,69	25,45	120,17
Axis 2	852,47	850,24	37,47	-2,23	776,2	924,28	769,77	922,16
Axis 3	59,72	60,98	20,08	1,26	21,3	100,66	26,37	106,68
Petal.Length								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	336,88	336,4	7,79	-0,48	321,01	351,78	320,66	350,58
Axis 2	0,6	1,1	1,2	0,5	-1,27	3,47	0	4,09
Axis 3	20,2	21,4	9,08	1,2	3,46	39,33	7,08	40,68
Petal.Width								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	319,06	318,66	8,03	-0,4	302,8	334,52	302,88	333,19
Axis 2	4,48	5,51	4,66	1,03	-3,7	14,72	0,12	18,17
Axis 3	402,3	400,92	25,46	-1,38	350,62	451,22	349,84	448,99

Tabla 14. Resultados para CREjFq datos iris

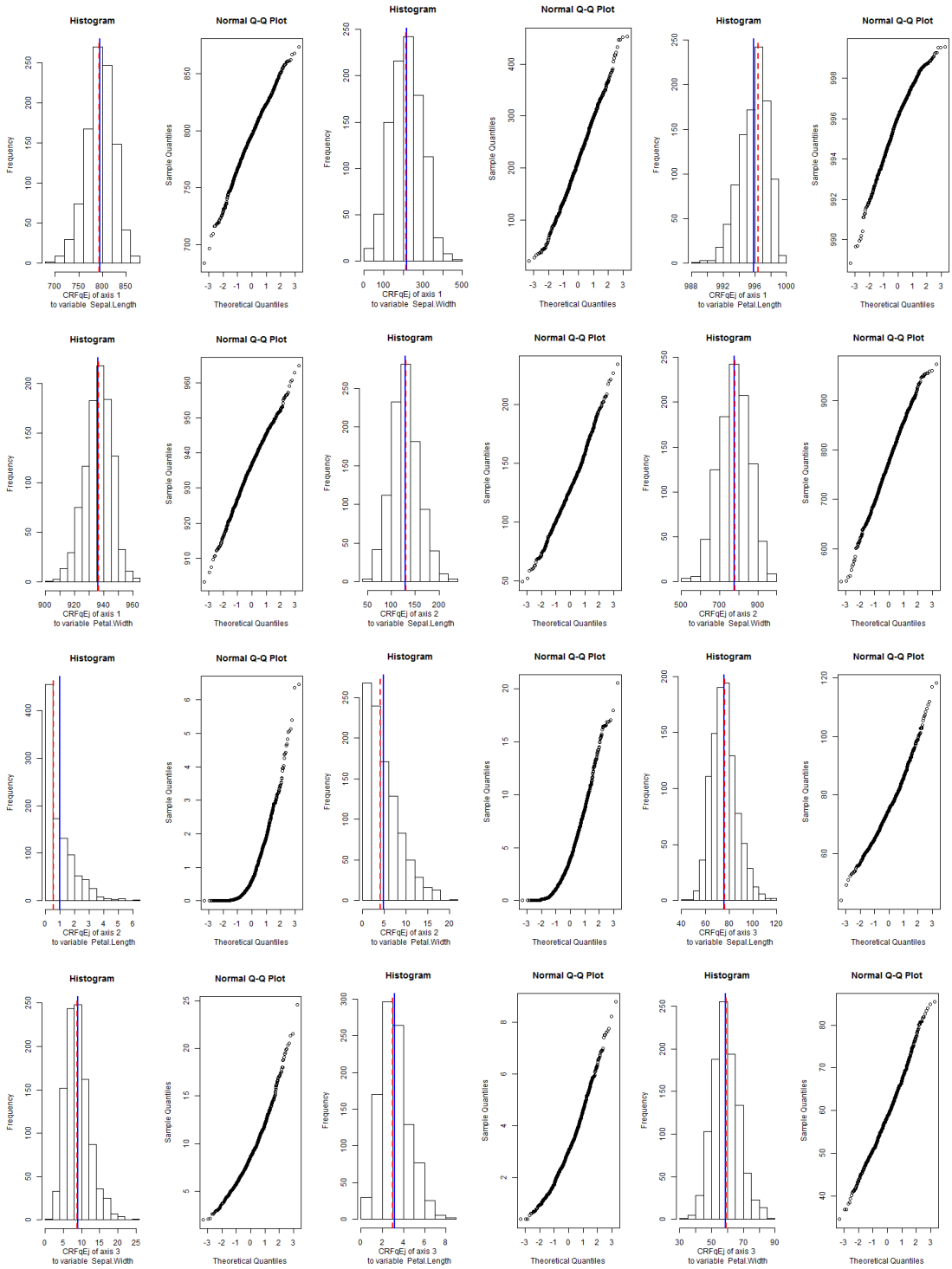


Figura 16. Histograma CRFqEj datos iris

Relative contribution of the factor q-th to column element j:								
Axis 1								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Sepal.Length	793,52	795,03	29,32	1,5	737,1	852,96	731,22	850,46
Sepal.Width	211,8	215,47	77,77	3,67	61,8	369,15	71,36	365,04
Petal.Length	996,44	995,87	1,78	-0,57	992,35	999,39	992,01	998,71
Petal.Width	936,5	936,24	9,24	-0,26	917,98	954,5	916,52	951,87
Axis 2								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Sepal.Length	130,38	129,29	29,76	-1,09	70,48	188,1	74,38	193,97
Sepal.Width	779,43	775,58	78,62	-3,85	620,23	930,93	623,57	923,17
Petal.Length	0,56	0,96	1,01	0,41	-1,04	2,97	0	3,44
Petal.Width	4,12	4,82	3,83	0,7	-2,75	12,38	0,12	14,24
Axis 3								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Sepal.Length	76,09	75,68	10,61	-0,41	54,71	96,65	57,48	98,74
Sepal.Width	8,77	8,95	3,29	0,18	2,44	15,45	3,73	16,96
Petal.Length	3	3,17	1,4	0,16	0,41	5,93	0,96	6,26
Petal.Width	59,38	58,94	8,32	-0,43	42,51	75,37	43,93	76,69

Tabla 15. Resultados para CRFqEj datos iris

Por último se muestra el gráfico con las coordenadas de las variables para todas las réplicas bootstrap (Figura 17). En él se han representado las variables con puntos con el objetivo de visualizar mejor los resultados. Se observa que la variable que tiene una variación mayor es la longitud del pétalo (Petal.Length).

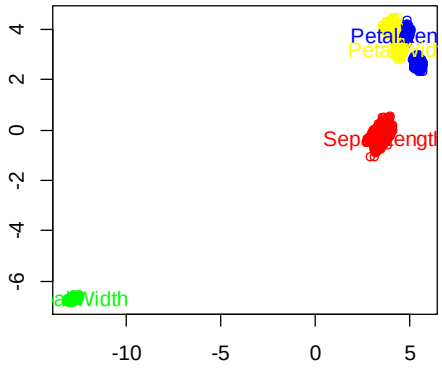


Figura 17. Coordenadas de las variables para las réplicas bootstrap datos iris

Datos Simulados

Los resultados para los datos simulados se muestran a continuación. En primer lugar se realizó un análisis HJ-Biplot con la transformación “Centrar por columnas”.

En la Tabla 16 se observa la información proporcionada por los valores propios. En ella se aprecia que el primer eje absorbe la mayor parte de la información (más del 50%) y con los tres primeros ejes se explica casi el 100% de la información (94,48%).

Valor Propio	Variabilidad Explicada	Acumulada
16,06	53,47	53,47
12,15	30,59	84,06
7,01	10,21	94,27

Tabla 16.Valores propios y variabilidad explicada datos simulados

La siguiente tabla (Tabla 17) recoge las contribuciones relativas del factor al elemento de las diferentes variables que se han analizado en los tres ejes que se han retenido.

	Eje 1	Eje 2	Eje 3
v1	226,48	16,48	737,49
v2	282,43	118,64	201,18
v3	281,58	89,64	46,09
v4	112,91	421,49	9,12
v5	96,6	353,75	6,12

Tabla 17. Contribuciones relativas del factor al elemento para las variables datos simulados

Según se puede apreciar, todas las variables están bien representadas en el primer plano principal (plano 1-2) excepto v1 que está bien representada en la tercera dimensión.

La representación HJ-Biplot en el primer plano principal se muestra en la Figura 18.

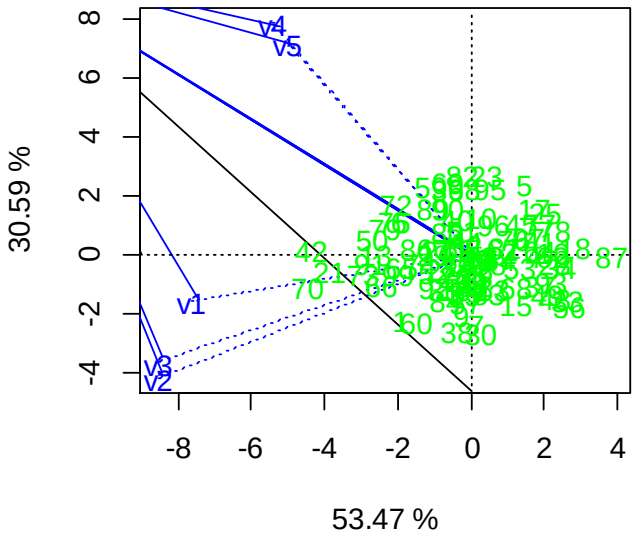


Figura 18. Representación HJ-BIPLOT de los datos simulados

La estructura de covariación de las variables pone de manifiesto una alta correlación entre las variables v4 y v5 y v2 y v3 representadas por un ángulo muy pequeño. Estas últimas variables tienen una correlación alta con la variable v1. Sin embargo, no tienen prácticamente relación con las variables v4 y v5 al presentar ángulos próximos a 90º.

Para analizar con más detalle estas relaciones, se muestran los ángulos entre las variables y entre las variables y los ejes en el plano 1-2 (Tabla 18 y Tabla 19).

	v1	v2	v3	v4	v5
v1	0	14,59	11,58	67,15	66,89
v2	14,59	0	3	81,73	81,48
v3	11,58	3	0	78,73	78,47
v4	67,15	81,73	78,73	0	0,26
v5	66,89	81,48	78,47	0,26	0

Tabla 18. Ángulos entre variables datos simulados

	Eje 1	Eje 2
v1	11,53	78,47
v2	26,12	63,88
v3	23,11	66,89
v4	55,62	34,38
v5	55,36	34,64

Tabla 19. Ángulos entre variables y ejes datos simulados

A continuación se muestran los resultados obtenidos aplicando el método a datos obtenidos mediante simulación.

Los parámetros estimados son los mismos que en el caso anterior.

En primer lugar se observa para la calidad de aproximación un valor esperado y calculado muy próximos y una concordancia entre los intervalos de confianza (Figura 19 y Tabla 20).

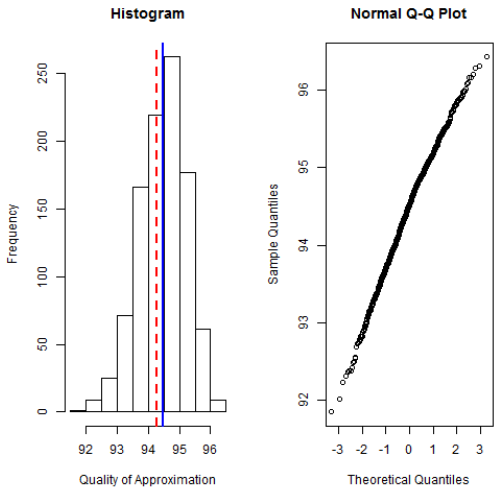


Figura 19.Histograma para calidad de aproximación datos simulados

Quality of approximation for columns:								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Quality of Approximation	94,27	94,46	0,75	0,19	92,98	95,95	92,9	95,8

Tabla 20. Resultados para calidad de aproximación datos simulados

El siguiente parámetro analizado son los valores propios. En los resultados se observa también poca variación entre valores observados y calculados y entre los dos tipos de intervalos calculados (Figura 20 y Tabla 21).

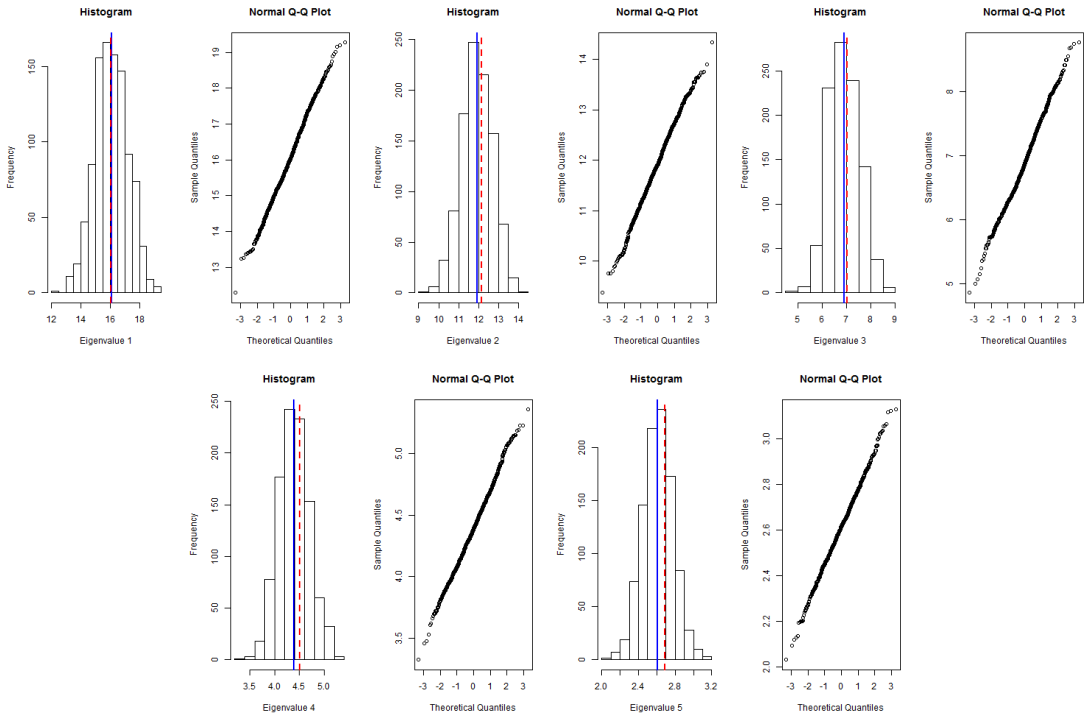


Figura 20. Histograma para valores propios datos simulados

Eigen values:								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Eigenvalue1	16,06	16,09	1,13	0,03	13,85	18,33	13,87	18,24
Eigenvalue2	12,15	11,92	0,78	-0,22	10,37	13,47	10,27	13,33
Eigenvalue3	7,01	6,9	0,63	-0,12	5,64	8,15	5,74	8,13
Eigenvalue4	4,51	4,39	0,31	-0,12	3,78	5	3,82	5,05
Eigenvalue5	2,69	2,61	0,17	-0,08	2,28	2,94	2,28	2,93

Tabla 21. Resultados para valores propios datos simulados

Para los casos de ángulos entre variables y ángulos entre variables y ejes (Figura 21, Figura 22, Tabla 22 y Tabla 23) las apreciaciones son las mismas que para el conjunto de datos anterior, es decir, hay poca variación entre valores observados y resultantes de la media de los valores calculados a partir de las muestras bootstrap y los intervalos t-bootstrap no son adecuados para aquellos ángulos próximos a cero.

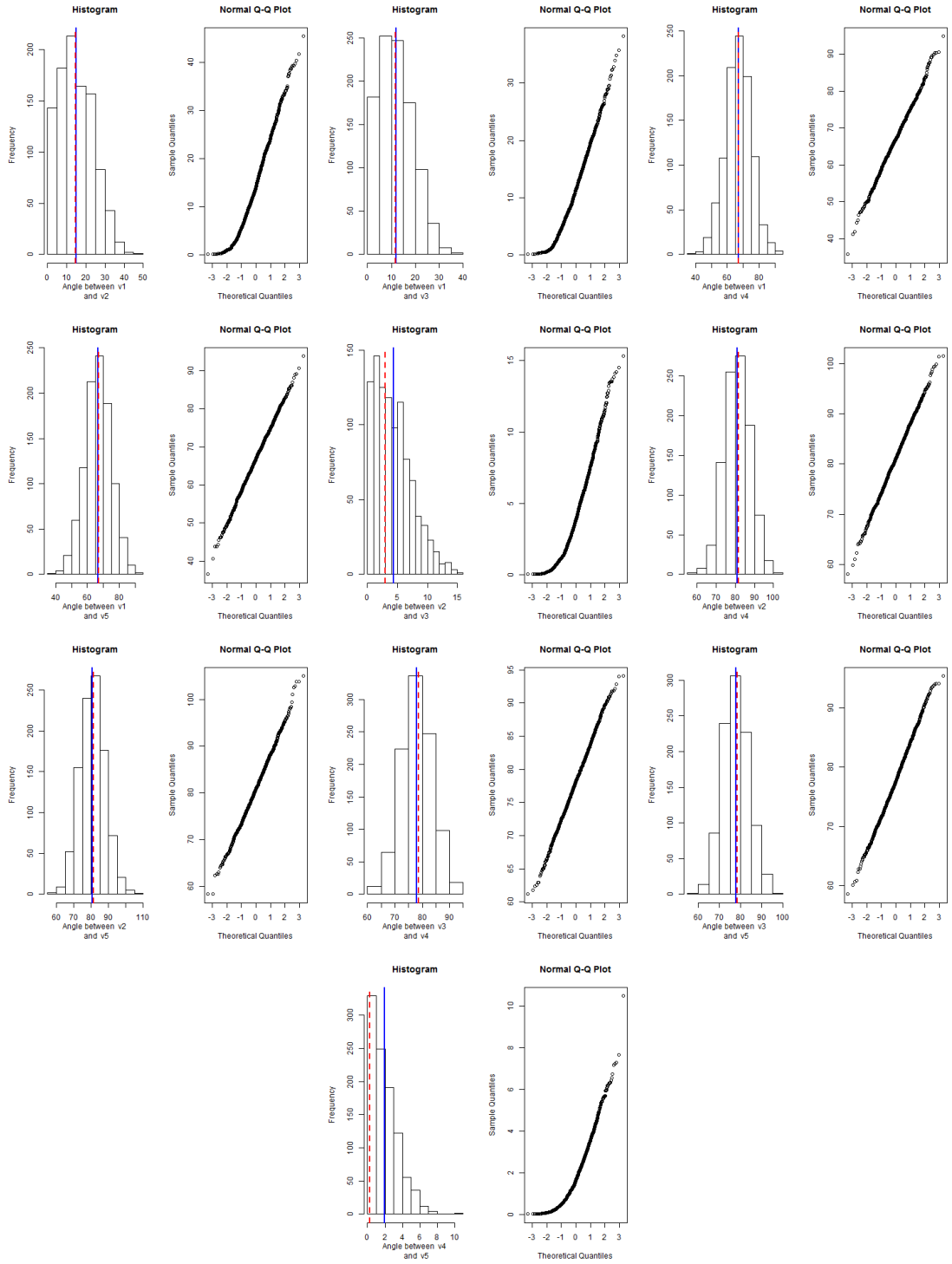


Figura 21. Histograma para ángulos entre variables datos simulados

Angles between variables:								
Angles between v1 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
v2	14,59	14,97	8,91	0,38	-2,72	32,65	0,95	33,37
v3	11,58	11,93	7,09	0,35	-2,14	25,99	0,65	26,44
v4	67,15	66,9	8,28	-0,25	50,46	83,33	50,09	83,05
v5	66,89	66,63	8,31	-0,26	50,15	83,11	49,83	82,39
Angles between v2 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
v3	3	4,35	3,09	1,35	-1,79	10,49	0,21	11,36
v4	81,73	81,06	6,79	-0,67	67,6	94,53	67,7	93,97
v5	81,48	80,8	7,26	-0,68	66,39	95,2	66,94	95,08
Angles between v3 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
v4	78,73	78,02	5,71	-0,71	66,7	89,34	66,78	89,55
v5	78,47	77,75	6,21	-0,72	65,43	90,08	65,98	90,36
Angles between v4 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
v5	0,26	1,98	1,55	1,72	-1,09	5,05	0,08	5,65

Tabla 22. Resultados para ángulos entre variables datos simulados

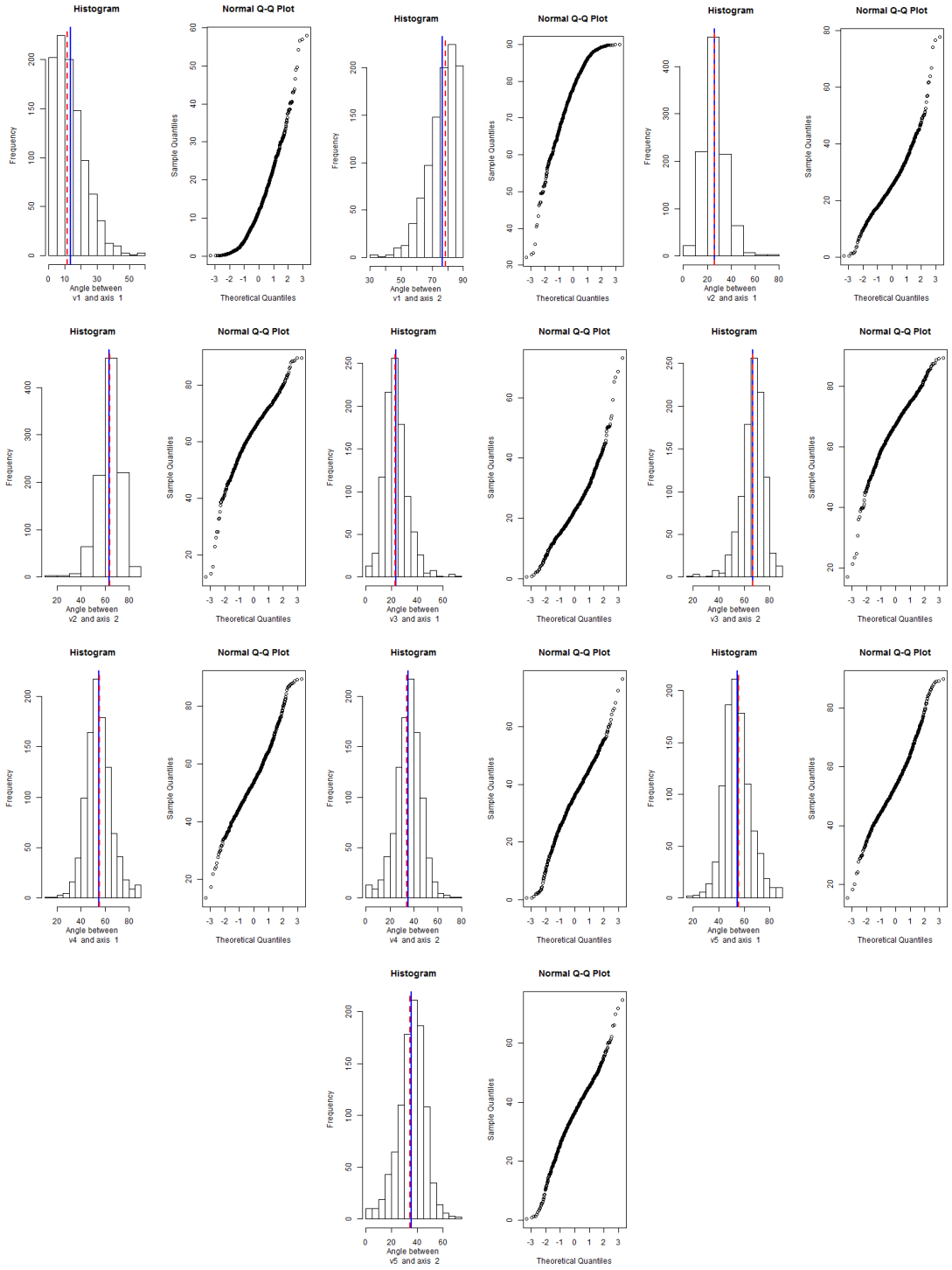


Figura 22. Histograma para ángulos entre variables y ejes datos simulados

Angles between variables and axes:								
Angles between v1 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	11,53	13,57	9,69	2,04	-5,66	32,8	0,78	37,38
Axis 2	78,47	76,43	9,69	-2,04	57,2	95,66	52,62	89,22
Angles between v2 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	26,12	26,35	9,52	0,23	7,46	45,23	10,29	46,86
Axis 2	63,88	63,65	9,52	-0,23	44,77	82,54	43,14	79,71
Angles between v3 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	23,11	23,39	9,03	0,28	5,46	41,32	7,85	43,47
Axis 2	66,89	66,61	9,03	-0,28	48,68	84,54	46,53	82,15
Angles between v4 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	55,62	54,64	10,76	-0,98	33,29	75,98	34,79	78,46
Axis 2	34,38	35,36	10,76	0,98	14,02	56,71	11,54	55,21
Angles between v5 and								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	55,36	54,34	10,64	-1,02	33,23	75,45	35,3	78,56
Axis 2	34,64	35,66	10,64	1,02	14,55	56,77	11,44	54,7

Tabla 23. Resultados para ángulos entre variables y ejes datos simulados

Si se observan los resultados referentes a las contribuciones de las variables a la variabilidad total se puede ver nuevamente que los datos tienen concordancia (Figura 23 y Tabla 24).

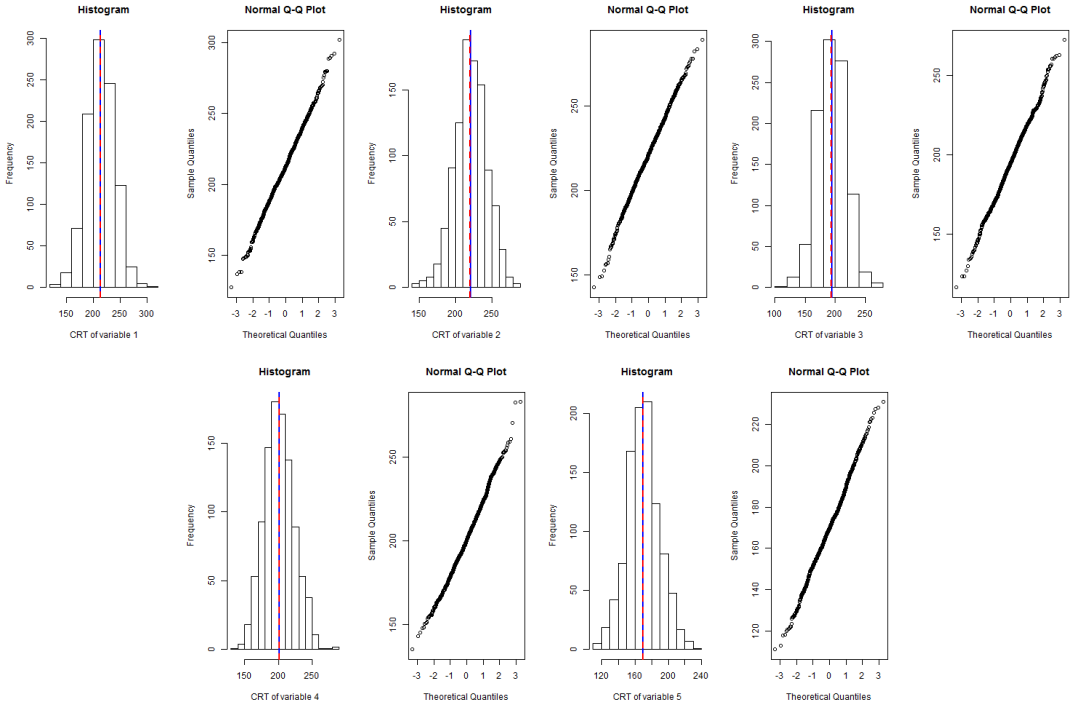


Figura 23. Histograma para contribuciones a la variabilidad total datos simulados

Relative contribution to total variability of the column element j:								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
v1	213,65	213,31	25,87	-0,33	161,98	264,65	162,5	264,24
v2	220,48	220,98	22,49	0,5	176,36	265,6	175,41	263,7
v3	193,8	194,43	23,89	0,63	147,02	241,84	146,89	239,9
v4	201,82	201,3	22,41	-0,52	156,83	245,76	160,78	246,76
v5	170,26	169,98	19,7	-0,28	130,89	209,07	130,4	209,98

Tabla 24. Resultados para contribuciones a la variabilidad total datos simulados

Por último, analizando los resultados obtenidos para las contribuciones relativas del factor al elemento y del elemento al factor (Figura 24, Figura 25, Tabla 25 y Tabla 26), se observan diferencias un poco mayores que en el resto de los parámetros y, al igual que en casos anteriores, se aprecia que los intervalos t-bootstrap no son adecuados para aquellos parámetros cuyas réplicas bootstrap presentan una asimetría mayor.

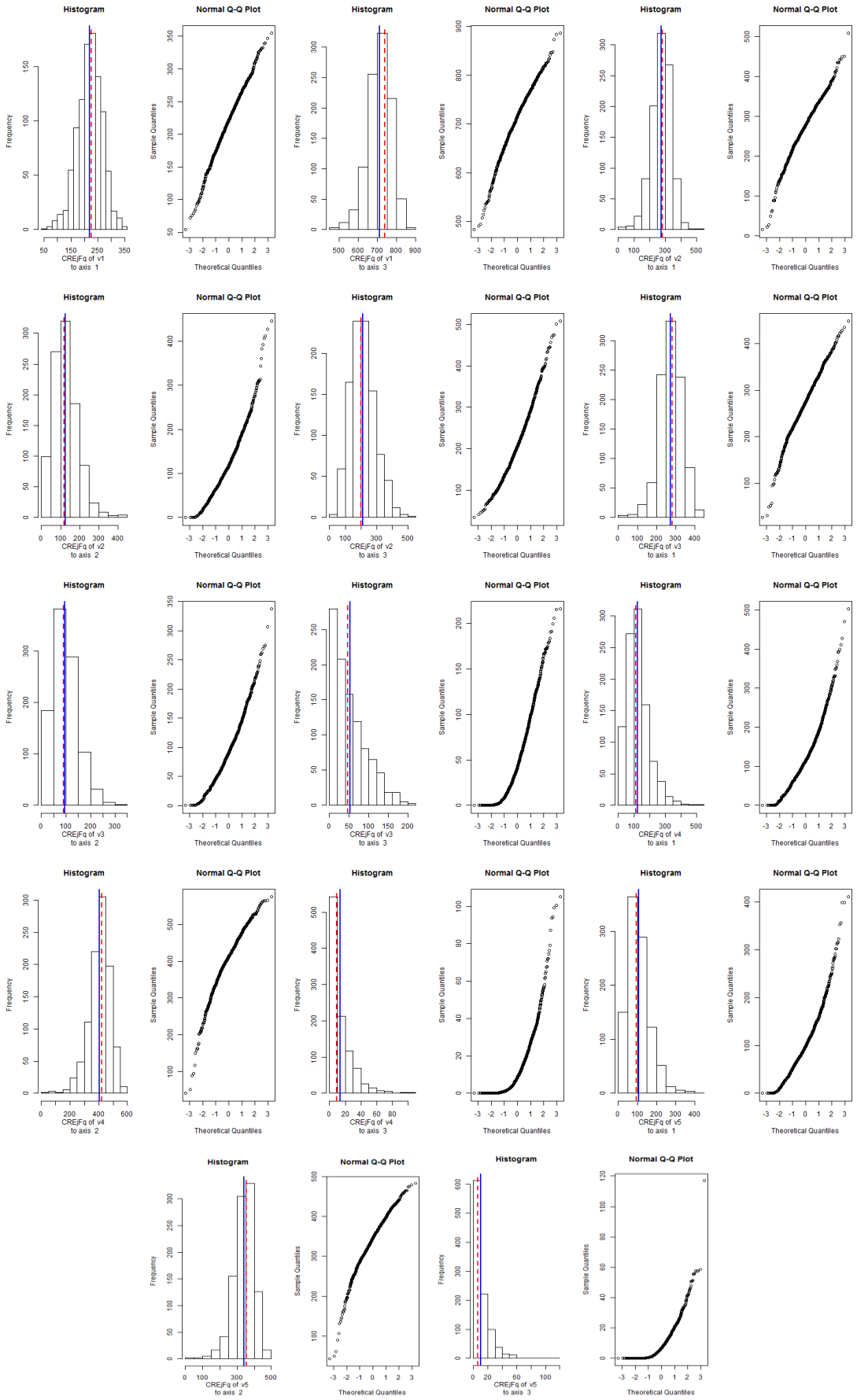
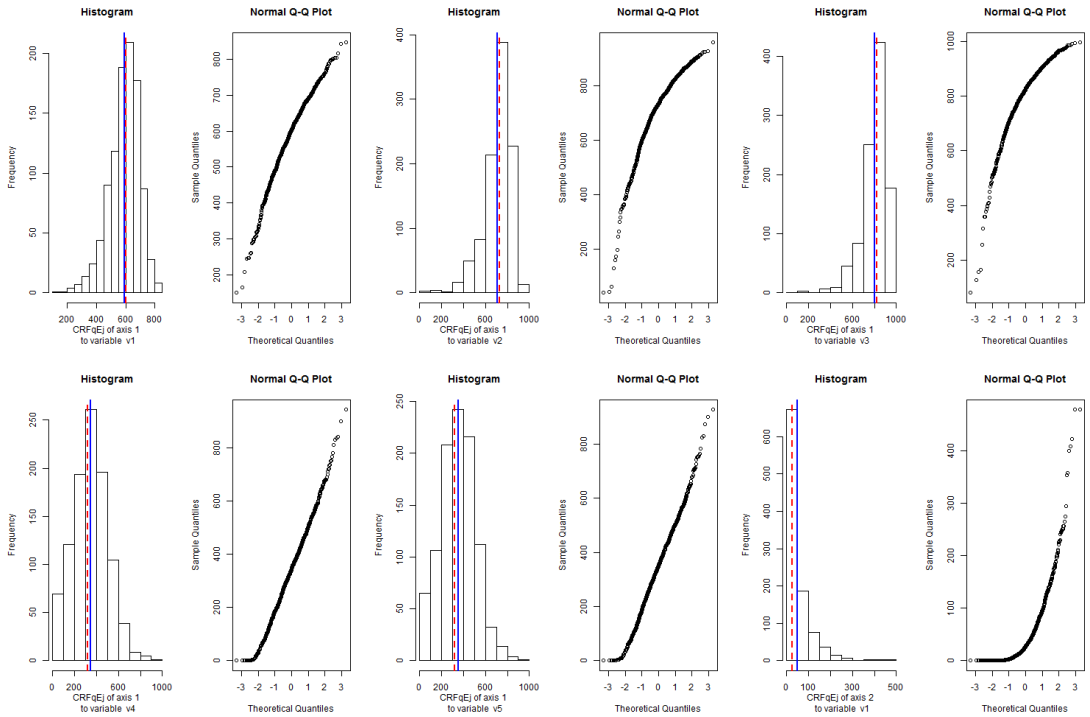


Figura 24. Histograma para CREjFq datos simulados

Relative contribution of the column element j to the factor q-th:								
v1								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	226,48	219,53	46,34	-6,95	127,58	311,48	119,52	308,53
Axis 2	16,48	31,97	38,24	15,49	-43,9	107,84	0,08	137,22
Axis 3	737,49	710,5	60,83	-26,99	589,79	831,21	570,42	813,29
v2								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	282,43	274,81	62,91	-7,62	149,97	399,65	141,98	388,04
Axis 2	118,64	126,5	65,48	7,86	-3,43	256,44	19,85	273,82
Axis 3	201,18	213,15	80,39	11,97	53,63	372,67	80,32	394,34
v3								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	281,58	273,8	60,05	-7,78	154,65	392,94	141,81	381,58
Axis 2	89,64	96,07	51,7	6,43	-6,5	198,65	11,78	211,87
Axis 3	46,09	52,41	43,66	6,33	-34,22	139,04	0,34	161,04
v4								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	112,91	125,13	72,24	12,22	-18,21	268,47	12,74	299,11
Axis 2	421,49	405,61	74,83	-15,88	257,14	554,09	232,81	527,45
Axis 3	9,12	13,6	15,5	4,47	-17,16	44,35	0,03	55,11
v5								

	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
Axis 1	96,6	106,73	60,14	10,13	-12,6	226,06	11,71	249,01
Axis 2	353,75	339,84	60,95	-13,9	218,9	460,79	197,73	439,52
Axis 3	6,12	10,35	11,44	4,23	-12,34	33,04	0,01	40,65

Tabla 25. Resultados para CREjFq datos simulados



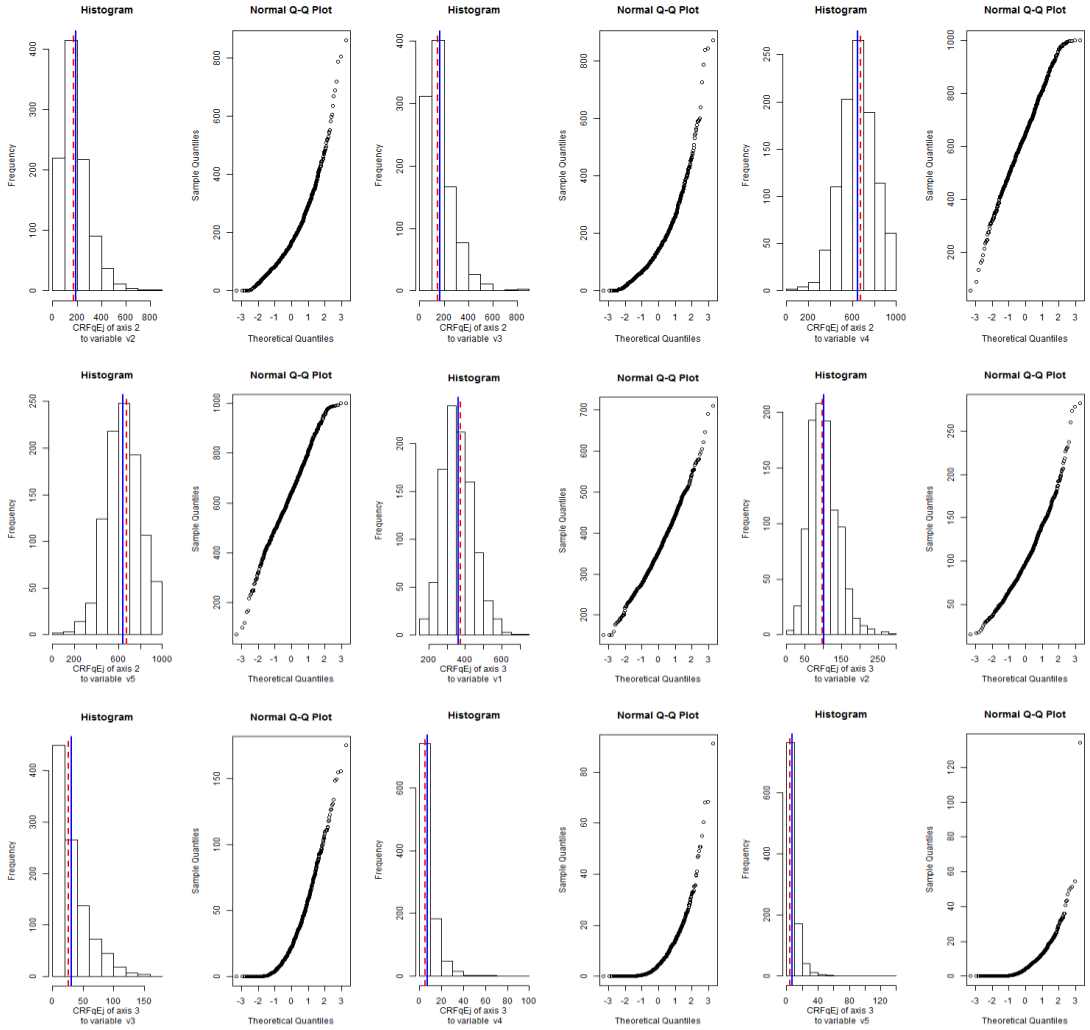


Figura 25. Histograma para CRFqEj datos simulados

Relative contribution of the factor q-th to column element j:								
Axis 1								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
v1	601,3	590,61	104,79	-10,69	382,69	798,54	336,57	757,85
v2	726,6	709,97	128,76	-16,62	454,49	965,46	391,59	884,17
v3	824,14	802,11	118,97	-22,03	566,05	1038,17	514,05	960,93
v4	317,35	344,85	159,49	27,5	28,38	661,31	39,53	672,06
v5	321,83	349,76	158,15	27,93	35,96	663,55	38,99	663,64

Axis 2								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
v1	25,03	48,68	61,38	23,66	-73,1	170,47	0,12	209,42
v2	174,63	188,02	116,01	13,39	-42,16	418,2	28,41	472,18
v3	150,12	166,92	116,59	16,81	-64,42	398,27	18,46	457,91
v4	677,76	647,79	159	-29,97	332,29	963,29	325,47	955,56
v5	674,28	643,44	157,02	-30,84	331,88	955,01	333,08	951,59
Axis 3								
	Obs. Value	Mean	SE	Bias	IC t-boot inf	IC t-boot sup	IC perc inf	IC perc sup
v1	373,67	360,71	82,97	-12,97	196,08	525,33	220,7	539
v2	98,77	102	39,25	3,23	24,12	179,89	37,48	192,42
v3	25,74	30,97	28,93	5,22	-26,44	88,37	0,18	105,96
v4	4,89	7,36	9,14	2,47	-10,77	25,5	0,02	31,2
v5	3,89	6,8	8,76	2,91	-10,58	24,17	0,01	28,72

Tabla 26. Resultados para CRFqEj datos simulados

A continuación se muestran las coordenadas de las variables para todas las réplicas bootstrap (Figura 26).

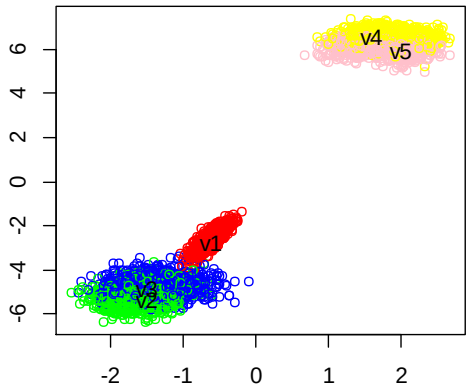


Figura 26. Coordenadas de las variables para las réplicas bootstrap datos simulados

Conclusiones

La metodología bootstrap permite ofrecer unos resultados completos del análisis HJ-Biplot. Mediante la combinación de ambas técnicas se proporciona una información más amplia ya que no sólo se muestran las estimaciones puntuales de las medidas que resultan del análisis HJ-Biplot, sino también indicadores de la precisión de estas estimaciones a través del cálculo de intervalos de confianza. Dichos intervalos de confianza se obtienen mediante los percentiles de la nueva muestra proporcionada por el remuestreo bootstrap y a través de los intervalos t-bootstrap. Para ello, se ha desarrollado un programa en el entorno R que permite obtener dichos resultados de una manera fácil y rápida a través de una interfaz gráfica de usuario amigable para el usuario y flexible a la hora de elegir parámetros a estimar y número de iteraciones y nivel de confianza para calcularlos.

Glosario

MCA: Análisis de Correspondencias Múltiple

PCA: Análisis de Componentes Principales

SVD: Descomposición en Valores Singulares

AMMI: Modelo de efectos principales aditivos e interacción multiplicativa

MLSCA: Análisis de Componentes Multinivel

GH-Biplot: Biplot que preserva la métrica para las columnas

JK-Biplot: Biplot que preserva la métrica para las filas

HJ-Biplot: Biplot que representa marcadores fila y columna en el mismo sistema de referencia con óptima calidad de representación

SQRT-Biplot: Biplot simétrico

Bibliografía

Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, *59*, 2–5.

Benzécri, J. P. (1973). *Analyse des Données*. Paris: Dunod.

Chatterjee, S. (1984). Variance estimation in factor analysis: An application of the bootstrap. *British Journal of Mathematical and Statistical Psychology*, *37*, 252–262.

Daudin, J., Duby, C., & Trécourt, P. (1988). Stability of principal components studied by the bootstrap method. *Statistics*, *19*, 241–258.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Diaconis, P., & Efron, B. (1983). Computer intensive methods in statistics. *Scientific American*, *248*, 116–130.

Eckart, G., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211–218.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171–185.

Efron, B., & Tibshirani, R. J. (1993). *An introduction into the bootstrap*. New York: Chapman and Hall.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, *Part II*, 179–188.

Gabriel, K. R. (1971). The Biplot graphic display of matrices with applications to principal components analysis. *Biometrika*, *58*(3), 453–467.

Galindo, M. P. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Qüestíio*, *10*(1), 13–23.

Galindo, M. P., & Cuadras, C. M. (1986). Una extensión del método Biplot y su relación con otras técnicas. *Publicaciones de Bioestadística y Biomatemática. nº 17. Universidad de Barcelona*.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: John Wiley & Sons.

Gower, J. (1992). Generalized biplots. *Biometrika, 79*(3), 475–493.

Gower, J., & Hand, D. (1996). *Biplots*. London: Chapman and Hall.

Gower, J., & Harding, S. (1988). Nonlinear biplots. *Biometrika*, (75), 445–455.

Greenacre, M. J. (1984). *Theory and Application of Correspondence Analysis*. London: Academic Press.

Holmes, S. (1985). *Outils Informatiques pour l'évaluation de la pertinence d'un résultat en analyse des données*. USTL, Montpellier, Francia.

Holmes, S. (1989). *Using the bootstrap and the RV coefficient in the multivariate context*. New York: Data Analysis, Learning Symbolic and Numeric Knowledge, E. Diday (ed.), Nova Science.

Jambu, M. (1991). *Exploratory and Multivariate Data Analysis*. Orlando: Academic Press, Inc.

Kiers, H. A. L. (2004). Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics, 18*, 22–36.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1990). Assessing sampling variation relative to number-of-factors criteria. *Educational and Psychological Measurement, 50*, 33–48.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research, 26*, 421–434.

Lavoranti, O. J., dos Santos, C. T., & Kraznowski, W. J. (2007). Phenotypic stability via ammi model with bootstrap re-sampling. *Pesq. Flor. bras.*, *54*, 45–52.

Lebart, L., Morineau, A., & Piron, M. P. (1995). *Statistique Exploratoire Multidimensionnelle*. Paris: Dunod.

Linting, M., Meulman, J. J., Groenen, P. J. F., & Van der Kooij, A. J. (2007). Stability of nonlinear principal components analysis. An empirical study using the balanced bootstrap. *Psychological Methods, 12*(3), 359–379.

Meulman, J. J. (1982). *Homogeneity Analysis of Incomplete Data*. Leiden: DSWO Press.

Milan, L., & Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, *44*, 31–49.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Soc. Series B*, *11*, 18–84.

R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raykov, T., & Little, T. D. (1999). A note on Procrustean rotation in exploratory factor analysis: A computer intensive approach to goodness-of-fit evaluation. *Educational and Psychological Measurement*, *59*, 47–57.

Stauffer, D. F., Garton, E. O., & Steinhorst, R. K. (1985). A comparison of principal component from real and random data. *Ecology*, *66*, 1693–1698.

Timmerman, M. E., Kiers, H. A. L., Smilde, A. K., Ceulemans, E., & Stouten, J. (2009). Bootstrap confidence intervals in multi-level simultaneous component analysis. *Br J Math Stat Psychol*, *62(Pt 2)*, 299–318.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

Van Ginkel, J. R., & Kiers, H. A. L. (2011). Constructing bootstrap confidence intervals for principal component loadings in the presence of missing data: a multiple-imputation approach. *Br J Math Stat Psychol*, *64*(3), 498–515.

Young, G., & Householder, A. S. (1938). Discussion of a Set of Points in Terms of Their Mutual Distances. *Psychometrika*, *3*, 19–22.